ORSAY
n° d'ordre :

UNIVERSITE DE PARIS-SUD

CENTRE D'ORSAY

THESE

présentée

pour obtenir

le Titre de Docteur 3ème Cycle
Spécialité : INFORMATIQUE

par

Danièle GARDY

SUJET : *EVALUATION DE RESULTATS D'OPERATIONS DE L'ALGEBRE RELATIONNELLE*

soutenue le *16 FEVRIER 1983* devant la Commission d'examen

MM.	N. SPYRATOS	Président
	R. DEMOLOMBE	
	E. GELENBE	Rapporteur
	M. METIVIER	
	C. CHOFFRUT	Invité

LIBRARY

ORSAY
n° d'ordre :

UNIVERSITE DE PARIS-SUD

CENTRE D'ORSAY

THESE

présentée

pour obtenir

le Titre de Docteur 3ème Cycle
Spécialité : INFORMATIQUE

par

Danièle GARDY

SUJET : *EVALUATION DE RESULTATS D'OPERATIONS DE L'ALGEBRE RELATIONNELLE*

soutenue le *16 FEVRIER 1983* devant la Commission d'examen

MM.	N. SPYRATOS	Président
	R. DEMOLOMBE	
	E. GELENBE	Rapporteur
	M. METIVIER	
	C. CHOFFRUT	Invité

Je remercie Monsieur E. GELENBE pour avoir accepté de diriger cette thèse, Monsieur N. SPYRATOS en tant que Président et Messieurs M. METIVIER, R. DEMOLOMBE et C. CHOFFRUT pour avoir participé au jury de thèse.

Je tiens à exprimer toute ma reconnaissance à Monsieur C. PUECH qui m'a suggéré l'emploi des fonctions génératrices pour ce travail.

Je remercie le Laboratoire de Mathématiques Appliquées de l'Ecole Polytechnique pour m'avoir accueillie pendant ma thèse.

Enfin, je remercie particulièrement Mesdames J. BAILLEUL et M-T. BOUVIER pour avoir assuré la frappe de cette thèse.

Plusieurs résultats de cette thèse (intersection, projection, jointures) ont été obtenus indépendamment par C. PUECH et doivent faire partie de sa thèse de doctorat ès-Sciences. Les méthodes de séries génératrices tirent leur origine de travaux de C. PUECH (février 1982).

TABLE DES MATIERES

Pages

RÉSUMÉ 9

CHAPITRE I - MODÈLE ET HYPOTHESES 11

A - INTRODUCTION 13

B - MODÈLE RELATIONNEL 15

C - MODÈLE PROBABILISTE 20

CHAPITRE II - SÉLECTION 23

A - INTRODUCTION 25

B - RELATION SANS DÉPENDANCE FONCTIONNELLE 27

I) LOIS UNIFORMES

II) LOIS GENERALES

C - RELATION AVEC UNE DÉPENDANCE FONCTIONNELLE SIMPLE 36

I) LOI DE $|s(R)| / |R|$

II) LOIS UNIFORMES

III) LOIS GENERALES (CAS OU $\Delta = Da \times Db$)

CHAPITRE III - UNION, INTERSECTION, DIFFÉRENCE 47

A - RELATION SANS DÉPENDANCE FONCTIONNELLE 50

I) LOIS UNIFORMES

II) LOIS GENERALES

B - RELATION AVEC UNE DÉPENDANCE FONCTIONNELLE SIMPLE 66

I) LOIS UNIFORMES

II) LOIS GENERALES

CHAPITRE IV - PROJECTION 79

A - RELATION SANS DÉPENDANCE FONCTIONNELLE 81

I) LOIS UNIFORMES
II) LOIS GENERALES

B - RELATION AVEC UNE DÉPENDANCE FONCTIONNELLE SIMPLE 98

I) LOIS UNIFORMES
II) LOIS GENERALES

CHAPITRE V - ÉQUIJOINTURE 107

A - LOIS UNIFORMES 110
B - LOIS GÉNÉRALES 131

CHAPITRE VI - DEMIJOINTURE 151

A - LOIS UNIFORMES 153
B - LOIS GÉNÉRALES 164

CONCLUSION 171

ANNEXE 175

APPENDICE 183

BIBLIOGRAPHIE 191

RÉSUMÉ

Nous cherchons dans ce travail à déterminer les tailles de relations intermédiaires obtenues lors de l'exécution d'une requête de l'algèbre relationnelle, soit la projection ou sélection d'une relation, ou l'union, intersection, différence ou équijoin de deux relations. Ces tailles sont utiles, par exemple, pour choisir entre deux écritures d'une requête celle qui est la plus avantageuse en temps, ou coût, ou bien pour réserver la place nécessaire en mémoire aux résultats intermédiaires.

Nous supposons connues des lois de probabilité sur les domaines des attributs intervenant dans les différentes relations de la base. Nous en déduisons les lois de probabilité des tailles des relations initiales et des relations obtenues par application d'opérations de l'algèbre relationnelle. Ceci nous permet de donner la loi conditionnelle de la taille d'une relation résultat, en supposant connues les tailles des relations initiales. Le cas de lois uniformes nous a paru mériter une étude plus approfondie : en effet, ces lois sont assez souvent rencontrées en pratique, et donnent des résultats plus simples que dans le cas général. Nous avons étudié deux types de relations : relations sans dépendances fonctionnelles (schémas libres) et relations avec une clé.

SUMMARY

We have given a probabilistic solution to the problem of estimating the size of intermediate results which appear during the evaluation of a query over a relational data base. We start from probability laws over the domains of attributes appearing in the data base and deduce probability laws over the values of given relations. We then proceed to give the generating function of the size of relations obtained by application of relational algebra operators (union or intersection, selection, projection, equijoin and semijoin) to these initial relations.

CHAPITRE I

MODELE ET HYPOTHESES

A - INTRODUCTION

Depuis sa présentation par Codd en 1970 [5], le modèle relationnel a permis de résoudre de manière élégante de nombreux problèmes posés par les bases de données : par exemple, indépendance entre le point de vue de l'utilisateur et l'implémentation physique, permettant l'emploi de langages d'interrogation de haut niveau, ou utilisation des formes normales et des contraintes d'intégrite pour éviter certaines incohérences.

Cependant, le fonctionnement de tels systèmes, de par le haut degré d'indépendance qu'ils offrent, pose de grands problèmes d'optimisation des requêtes. Ainsi, Ullman[17] a calculé que le produit cartésien de deux relations de taille 10.000 demendait, avec les hypothèses retenues, 35 minutes !

Des travaux d'optimisation ont donc été entrepris très tôt : dès 1975, Chang et Smith [4] ont étudié la décomposition d'une requête en plusieurs parties, pour permettre un plus grand parallélisme de traitement et éviter de calculer des expressions redondantes. Hall [11] a étudié la transformation d'une requête mal posée en une requête équivalente d'exécution plus simple. Aho, Sagiv et Ullman [1] se sont intéressés à l'équivalence de requêtes comprenant les opérateurs de jointure, projection et sélection et Yao [18] a présenté pour ces mêmes types de questions, des équations de coût permettant une comparaison plus fine des différentes solutions.

Demolombe [3] utilise de manière systématique de telles fonctions de coût pour comparer différentes formulations d'une question. Ces fonctions prennent en compte le temps CPU, la taille-mémoire et le

nombre d'entrées-sorties nécessaires, ce qui sous-entend la possibilité de connaître, ou au moins de pouvoir estimer, la taille des relations considérées. Ce dernier problème a été abordé brièvement par Rosenthal [16], qui a calculé la taille moyenne d'un join sous certaines hypothèses et surtout par Demolombe et Richard.

Demolombe [7], [8], [9] estime le nombre de n-uplets satisfaisant certaines formules du calcul relationnel.

Richard [14], [15] définit des systèmes de paramètres permettant d'évaluer les tailles des relations, résultats d'opérations de l'algèbre relationnelle et montre qu'on peut propager ces paramètres aux relations dérivées.

Nous avons entrepris [12], [13] une étude systématique dans le cadre de l'algèbre relationnelle. Nous définissons d'abord le modèle probabiliste retenu, qui diffère de ceux de Demolombe et Richard essentiellement par une plus grande simplicité et un moindre nombre de paramètres (ce modèle nous permet d'ailleurs de calculer les propres paramètres de Richard). Nous calculons ensuite les tailles des résultats d'opérations ensemblistes : union, intersection, différence (Ch. III), de sélections (Ch. II), de projections (Ch. IV), d'équijointure (Ch. V) et de demi-jointure (Ch. VI).

B - MODÈLE RELATIONNEL

Le modèle de base de données retenu est le modèle relationnel, où les données sont rangées dans des tableaux.

Nous définissons une *relation* par un *schéma* de relation : R[A1 : D1,...,An : Dn] , où les A_i sont les *attributs* de la relation, et prennent leurs valeurs dans les *domaines* D_i . Une *valeur* de la relation est un sous-ensemble du produit cartésien $D1 \times ... \times Dn$.

Dans la suite, nous confondrons souvent schéma et valeur de relation, appelés tous deux : "relation", lorsqu'il n'y a pas d'ambiguité.

Contraintes d'intégrité

Une représentation adéquate du monde réel introduit souvent des restrictions sur l'ensemble des valeurs qu'une relation peut prendre.

Par exemple, personne n'est âgé de - 13 ou 200 ans, ou n'habite le sommet de l'Everest. Ce genre de contrainte est souvent résolu par un choix convenable des domaines des attributs. D'autres restrictions, comme le fait qu'un individu donné a au plus un numéro de Sécurité Sociale, se traduisent par des *dépendances fonctionnelles* :

Si A et B sont deux sous-ensembles des attributs d'une relation R , B dépend fonctionnellement de A (noté $A \rightarrow B$), si : pour toute valeur r de la relation, si (a,b) est une partie d'un n-uplet de r , et (a,b') une partie d'un autre n-uplet de r , alors : b = b'.

Dans la relation Famille [Père, Mère, Enfant] , nous avons les dépendances fonctionnelles suivantes : $\begin{cases} \text{Enfant} \rightarrow \text{Père} \\ \text{Enfant} \rightarrow \text{Mère} \end{cases}$

mais les dépendances fonctionnelles Père → Enfant et (Père, Mère) → Enfant sont fausses.

Nous considérons ici deux types de schémas :

1) Schémas "libres" :

R[A1 : D1 ,..., An : Dn] , sans dépendance fonctionnelle ;

2) Schémas "à clé" :

Nous supposons qu'il existe une clé, ie. un sous-ensemble X des attributs $\{A_1,...A_n\}$ de R , tel que : $\forall$ i , $1 \leqslant i \leqslant n$, $X \rightarrow A_i$.

Nous supposons de plus qu'il n'y a pas d'autres dépendances fonctionnelles que celles qui se déduisent des précédentes.

Nous pouvons donc nous ramener au cas d'une relation à deux attributs A et B , non nécessairement simples, où $A \rightarrow B$, en posant :

$A = X$

$B = \{A_1 \ldots A_n\} \setminus X$.

Les domaines correspondants sont :

$$D_A = \prod_{A_i \in X} D_i \ .$$

$$D_B = \prod_{A_i \notin X} D_i \ .$$

Algèbre relationnelle

Le langage d'interrogation sur la base de données est l'algèbre relationnelle :

Toute interrogation de la base de donnée se traduit par une suite d'opérations élémentaires de l'algèbre sur des relations de la base, et fournit une relation-résultat.

Les opérations de base de l'algèbre relationnelle sont (cf. [Ullmann]) :

- *l'union* et la *différence* de deux relations de même arité. Ces opérations sont définies au sens ensembliste. Nous les appliquerons ici à des valeurs distinctes d'un même schéma de relation.

- le *produit cartésien* de deux relations : si R a ℓ attributs, et S, m attributs, le produit cartésien de R est l'ensemble des $(\ell + m)$-uplets dont les ℓ premiers éléments forment un ℓ-uplet de R, et les m suivants un m-uplet de S.

- la *projection* sur un sous-ensemble $\{A_{i_1},\ldots,A_{i_p}\}$ des attributs $\{A_1,\ldots,A_n\}$ de R : notée $\Pi_{i_1\ldots i_p}(R)$, c'est l'ensemble des p-uplets $(x_{i_1},\ldots x_{i_p})$, qui vérifient : $\exists\ (a_1,\ldots a_n) \in R$, tel que : $\forall\ j \in \{1,\ldots,p\}$, $x_{i_j} = a_{i_j}$.
  Cela revient à supprimer les attributs non dans la projection, puis à éliminer les doubles de la relation obtenue.

- la *sélection* :

  Soit C une condition, comprenant :

  - des valeurs ou noms d'attributs comme opérantes
  - des opérateurs arithmétiques de comparaison : $<$, $\leqslant$, $=$, $>$, $\geqslant$, $\neq$.
  - des opérateurs logiques : $\wedge$, $\vee$, $\neg$.

  La sélection de la relation R suivant C est l'ensemble des n-uplets de R, vérifiant C. Nous la notons $s_C(R)$, ou, lorsqu'il n'y a pas risque de confusion, $s(R)$.

  Nous mettrons en bijection une sélection associée à une condition C, et l'ensemble Δ de tous les n-uplets possibles qui vérifient C.

$$s(R) = R \cap \Delta \text{ , et } \Delta = \bigcup_{\text{valeurs de } R} s(R) .$$

Les opérateurs suivants, bien que pouvant s'écrire à l'aide des opérateur déjà énumérés, sont souvent utilisés :

. l'intersection : $R \cap S = R \setminus (R \setminus S)$

. le θ-join, où θ est un opérateur arithmétique de comparaison.

Soit A un attribut de R, B un attribut de S . Le θ-join de R sur A, et S sur B , est l'ensemble des n-uplets de $R \times S$, tels que : $A \ \theta \ B$.

Il est noté : $R \underset{A \theta B}{\times} S$

. l'équijoin : c'est le θ-join, dans le cas où θ est l'égalité, et où l'on ne garde qu'un des attributs sur lesquels porte la comparaison.

Par exemple, soient R[A,B] et S[A,C] deux relations, prenant les valeurs suivantes :

R :

A	B
a	b
a	b'
a'	b''

S :

A	C
a	c
a'	c
a''	c'

l'équijoin de R et S sur A , noté $R \underset{A}{\bowtie} S$, ou simplement $R \bowtie S$, est :

T[A,B,C] :

A	B	C
a	b	c
a	b'	c
a'	b''	c

. le semi-join de R par S est la projection sur R de l'équijoin $R \bowtie S$: $\Pi_R(R \bowtie S)$. L'intérêt de cette opération est qu'il est possible, dans certains

types de requêtes [2], de remplacer les joins par des semi-joins, de taille souvent nettement inférieure.

En effet : * $|R \times S| \leqslant |R| \times |S|$, et peut-être du même ordre de grandeur

* $|\Pi_R(R \times S| \leqslant |R|$.

C - MODÈLE PROBABILISTE

Nous cherchons à résoudre le problème suivant : "connaissant les tailles de deux relations R et S, quelle est la loi de probabilité de la taille de la relation T obtenue en appliquant à R et S, ou seulement à S, un opérateur de l'algèbre relationnelle ? En particulier, quelles sont la taille moyenne et la variance de T ? "

Pour cela, nous définissons d'abord, pour chaque schéma de relation, une loi de probabilité sur les valeurs possibles, telle que la probabilité d'obteni une valeur R, soit proportionnelle à la probabilité de chacun des éléments des n-uplets ("lignes") de R :

Nous supposons donné, pour chaque attribut A, son domaine D_A, et une loi de probabilité sur D_A. Ceci revient à définir une variable aléatoire X_A à valeurs dans D_A, et à noter : $p(a) :=$ Probabilité $(X_A = a)$.

La loi de probabilité sur D_A est indépendante des relations comprenant l'attribut A.

Le problème pratique qui peut se poser est l'estimation de cette loi de probabilité : souvent, elle n'est pas connue a priori, et on l'estime en regardant les occurrences d'une valeur a donnée dans différents états de la base.

Pour chaque schéma de relation, nous définissons ensuite une loi de probabilité sur l'ensemble des valeurs de cette relation :

$$P(R) = k. \prod_{\ell \in R} . \prod_{a_i \in \ell} p(a_i) \quad ,$$

où k est une constante de "normalisation", définie par : $\sum_{\text{valeurs } R} P(R) = 1$, et

dépendant du type de schéma considéré :

* schéma libre : $R[A_1 : D_1 , \ldots , A_n : D_n]$.

Soit $D = D_1 \times \ldots \times D_n = \{\ell_1 , \ldots , \ell_d\}$, et $p_i = p(\ell_i)$, $1 \leqslant i \leqslant d$.

Soit $R = \{\ell_{i_1} , \ldots , \ell_{i_r}\}$ une valeur de la relation considérée :

$$P(R) = k \cdot \prod_{j=1}^{r} p_{i_j} .$$

$$\sum_R P(R) = \sum_{r \geqslant 0} \sum_{i_1 \ldots i_r} k \cdot p_{i_1} \ldots p_{i_r}$$

$$= k \cdot \prod_{i=1}^{d} (1 + p_i) .$$

Donc : $$P(R) = \frac{\prod_{\ell \in R} p(\ell)}{\prod_{i=1}^{d} (1 + p_i)} .$$

* Schéma à clé : $R[A : D_A , B : D_B]$.

Notons $D_A = \{x_1 , \ldots , x_{d_A}\}$, et $p_i = p(x_i)$, $1 \leqslant i \leqslant d_A$

$D_B = \{y_1 , \ldots , y_{d_B}\}$, et $q_j = p(y_j)$, $1 \leqslant j \leqslant d_B$.

Soit $R = \{x_{i_1} y_{j_1} , \ldots , x_{i_r} y_{j_r}\}$ une valeur de la relation : A étant clé, les x_i sont tous distincts, mais non les y_j ,

$$P(R) = k \cdot \prod_{k=1}^{r} p_{i_k} \cdot q_{j_k} .$$

$$\sum_R P(R) = k \cdot \sum_{r \geqslant 0} \sum_{\substack{i_1 \ldots i_r \\ \text{(tous distincts)}}} p_{i_1} \ldots p_{i_r} \cdot \sum_{j_1 \ldots j_r} q_{j_1} \ldots q_{j_r}$$

$$\sum_{j_1 \ldots j_r} q_{j_1} \ldots q_{j_r} = (q_1 + \ldots + q_{d_B})^r = 1$$

Donc $$\sum_{R} P(R) = k \cdot \prod_{i=1}^{d_A} (1 + p_i)$$

et $$P(R) = \frac{\prod_{\ell \in R} p(\ell)}{\prod_{i=1}^{d_A} (1 + p_i)} .$$

Taille du résultat d'une opération

Soit f une application : $\{R\} \to \mathbb{R}$, ou $\{R\} \times \{S\} \to \mathbb{R}$. Nous avons défini des lois de probabilité, P_R, ou P_R et P_S, sur les ensembles de départ. Nous cherchons maintenant la loi-image $P_{(f)}$:

soit $\Delta \subset \mathbb{R}$, $P[\{R, \text{ tel que } f(R) \in \Delta\}] = P_{(f)}(\Delta)$.

Par exemple : . $f(R) = |s(R)|$, où s est une sélection ;

. $f(R) = |\Pi_A(R)|$, A attribut de R ;

. $f(R,S) = |R \times S|$ ...

Nous étudierons ce problème dans le cas où $f = |g|$, g étant un des opérateurs relationnels définis plus haut. Nous ferons une étude générale, pc des lois de probabilité quelconque, et donnerons aussi des résultats dans le cas particulier où les lois de probabilité sur les domaines sont uniformes.

CHAPITRE II

SELECTION

A - INTRODUCTION

Nous considérons une relation de l'un des types suivants :

* $R[A_1, \ldots, A_n]$, les attributs A_i ne satisfaisant à aucune dépendance fonctionnelle ;

* $R[A,B]$, où B dépend fonctionnellement de A (noté $A \rightarrow B$) .

Soit $D = \prod_{i=1}^{n} \mathrm{Dom}(A_i)$ (1er cas)

ou $\mathrm{Dom}(A) \times \mathrm{Dom}(B)$ (2e cas)

Nous définissons une sélection s sur la relation R , par un sous-ensemble Δ de D , associé : $s(R) = \{$n-uplets de R , qui sont dans $\Delta\}$.
Il y a donc bijection entre les sélections sur un schéma de relation donné, et les sous-ensembles du domaine D associé.

Nous obtenons la fonction génératrice de $|s(R)|$ en fonction de Δ , et en déduisons la taille moyenne et la variance de $|s(R)|$.

Le problème de la taille du résultat d'une sélection s passe donc par la détermination explicite de l'ensemble Δ associé. Ceci peut être simple (un cas couramment rencontré est : $\Delta = \prod_{i=1}^{n} \Delta_i$, $\Delta_i \subseteq D_i$: "$\Delta_i = D_i$" revient à ne pas imposer de condition à l'attribut Δ_i). La détermination de Δ peut aussi être assez compliquée, comme dans l'exemple suivant :

Nous effectuons sur la relation $R[A1 : D1, \ldots, An : Dn]$ une sélection s définie par : $\sum_{i=1}^{n} A_i = k$.

Pour $D_i \subseteq \mathbb{N}$, le problème posé se ramène à un problème de partition de l'entier k en n sommants $\geqslant 0$. La détermination de Δ se fait de manière

exhaustive. Or, pour de "grandes" valeurs de k , le nombre $\delta = |\Delta|$ de partitio
vaut à peu près $\frac{k^{n-1}}{n!(n-1)!}$, et devient rapidement très grand [6].

Taille moyenne

Dans le cas de lois uniformes, nous pouvons toujours calculer la taille moyenne de $s(R)$:

$$E[|s(R)|] = |R| . \frac{\delta}{d} \quad \text{où} \quad \delta = |\Delta| \quad d = |D|$$

En effet :

$$|s(R)| = \sum_{\ell \in \Delta} 1_{(\ell \in R)} .$$

Donc : $E[|s(R)|] = \sum_{\ell \in \Delta} P(\ell \in R)$

or : $P(\ell \in R) = \frac{|R|}{|D|}$.

D'où : $E[|s(R)|] = |R| . \frac{|\Delta|}{|D|}$. ■

B - RELATION SANS DÉPENDANCE FONCTIONNELLE

Nous supposons ici que la relation $R[A_1 \ldots A_n]$ n'a pas de dépendance fonctionnelle. Nous pouvons donc nous ramener à une relation $R[A]$ à un attribut (non simple) ; la sélection $s(R)$ est alors : $\{a \in R$, tel que $a \in \Delta\}$. Nous noterons dans la suite :

$$D = \prod_{i=1}^{n} D_i, \quad d = |D|$$

$$\Delta \subseteq D, \quad \delta = |\Delta| \ .$$

Un cas souvent rencontré en pratique est celui où $\Delta = \prod_{i=1}^{n} \Delta_i$, $\Delta_i \subseteq D_i$ ($\Delta_i = D_i$ revient à ne pas imposer de condition à l'attribut A_i).

I - LOIS UNIFORMES

La probabilité que $s(R)$ soit de taille t , sachant que R est de taille r , est :

$$P[|s(R)| = t \ / \ |R| = r] = \frac{\binom{d-\delta}{r-t}\binom{\delta}{t}}{\binom{d}{r}} \ .$$

La taille moyenne de $s(R)$ est :

$$E = r \cdot \frac{\delta}{d} \ ,$$

et la variance :

$$\sigma^2 = \frac{r\,\delta(d-r)\,(d-\delta)}{d^2(d-1)} \ .$$

Si $r \ll \delta$, notons $p = \frac{\delta}{d}$:

$$P[|s(R)| = t \;/\; |R| = r] \simeq \binom{r}{t} \cdot p^t \cdot (1-p)^{r-t}$$

$$E[|s(R)| \;/\; |R| = r] = r.p$$

$$\sigma^2 \simeq r.p.\left(1 - p - \frac{r-1}{d}\right) .$$

Démonstration

1) Les valeurs de R , de taille donnée r , étant équiprobables :

$$P[|s(R)| = t \;/\; |R| = r] = \frac{N}{N_r}$$

où N_r est le nombre de valeurs de R de taille r

N est le nombre de ces valeurs, qui vérifient de plus : $|s(R)| = t$.

. $N_r = \binom{d}{r}$

. $N = \binom{\delta}{t} \times \binom{d-\delta}{r-t}$:

en effet, pour construire R , nous choisissons d'abord t lignes dans Δ , puis complétons par $(r-t)$ lignes dans $D \setminus \Delta$.

Donc : $$P[|s(R)| = t \;/\; |R| = r] = \frac{\binom{d-\delta}{r-t}\binom{\delta}{t}}{\binom{d}{r}} .$$

2) Taille moyenne

$$|s(R)| = \sum_{\ell \in \Delta} 1_{(\ell \in R)}$$

donc : $$E[|s(R)|] = \sum_{\ell \in \Delta} E[1_{(\ell \in R)}]$$

$$= \sum_{\ell \in \Delta} P(\ell \in R)$$

Or, toutes les lignes ℓ étant équiprobables :

$$P(\ell \in R) = \frac{|R|}{d}$$

Donc : $$E[|s(R)|] = \frac{r \cdot \delta}{d} \quad .$$

3) Variance

$$\sigma^2 = E(|s(R)|^2) - [E(|s(R)|)]^2$$

$$E(|s(R)|^2) = \sum_{t=0}^{r} t^2 . P[|s(R)| = t \ / \ |R| = r]$$

$$= \sum_{t=1}^{r} t(t-1) . P[|s(R)| = t \ / \ |R| = r]$$

$$+ \sum_{t=1}^{r} t . P[|s(R)| = t \ / \ |R| = r]$$

$$\cdot \sum_{t=1}^{r} t.P[|s(R)| = t \;/\; |R| = r] = E[|s(R)| \;/\; |R| = r]$$

$$\cdot \sum_{t=1}^{r} t(t-1)\, P[|s(R)| = t \;/\; |R| = r]$$

$$= \frac{1}{\binom{d}{r}} \sum_{t=1}^{r} t(t-1).\binom{d-\delta}{r-t}\binom{\delta}{t}$$

$$= \frac{\delta(\delta-1)}{\binom{d}{r}} \sum_{t=2}^{r} \binom{d-\delta}{r-t}\binom{\delta-2}{t-2}$$

$$= \frac{\delta(\delta-1)}{\binom{d}{r}} \sum_{t=0}^{r-2} \binom{(d-2)-(\delta-2)}{(r-2)-t}\binom{\delta-2}{t}$$

Or : $$\sum_{t=0}^{r} \binom{d-\delta}{r-t}\binom{\delta}{t} = \binom{d}{r} \cdot$$

Donc : $$\sum_{t=1}^{r} t(t-1)\, P[|s(R)| = t \;/\; |R| = r] = \frac{\delta(\delta-1)}{\binom{d}{r}} \binom{d-2}{r-2}$$

$$= \frac{\delta(\delta-1)\; r(r-1)}{d(d-1)} \cdot$$

Nous en tirons :

$$\sigma^2 = \frac{\delta(\delta-1)\; r(r-1)}{d(d-1)} + \frac{r\delta}{d}\left(1 - \frac{r\delta}{d}\right)$$

$$\sigma^2 = \frac{r\delta}{d}\left[1 - \frac{r\delta}{d} + \frac{(r-1)(\delta-1)}{d-1}\right]$$

$$\sigma^2 = \frac{r\delta(d-r)\ (d-\delta)}{d^2(d-1)}$$

4) Approximations

Supposons $r \ll \delta$, d, et notons $p = \frac{\delta}{d}$.

En utilisant l'approximation :

$$\binom{d}{x} \sim \frac{d^x}{x!} \quad \text{pour} \quad x \ll d \ ,$$

nous obtenons :

$$\cdot\ P[|s(R)| = t \ / \ |R| = r] \simeq \frac{(d-\delta)^{r-t}\ \delta^t\ r!}{(r-t)!\ t!\ d^r}$$

$$\simeq \binom{r}{t}\ (1-p)^{r-t}\ .\ p^t$$

$$\cdot\ \sigma^2 = \frac{r\delta(d-2)(d-\delta)}{d^2(d-1)} = \frac{r\ p\left(1 - \frac{r}{d}\right)(1-p)}{\left(1 - \frac{1}{d}\right)}$$

$$\sigma^2 \simeq r\ p\left(1 - \frac{r-1}{d} - p\right) . \quad \blacksquare$$

II - LOIS GENERALES

La probabilité que $s(R)$ soit de taille t, sachant que R est de taille r, est :

$$P[|s(R)| = t \;/\; |R| = r] = \frac{[x^t y^r]\ \phi(x,y)}{[y^r]\ \varphi(y)}$$

où $$\phi(x,y) = \prod_{\ell=1}^{d} (1 + p_\ell . x^{1(x_\ell \in \Delta)}\ y)$$

et $$\varphi(y) = \phi(1,y) = \prod_{\ell=1}^{d} (1 + p_\ell\ y)\ .$$

La taille moyenne de $s(R)$ est :

$$E = P[\Delta] - \frac{[y^r]\ \left\{\varphi(y) . \sum_{\ell : x_\ell \in \Delta} \frac{1}{1 + p_\ell\ y}\right\}}{[y^r]\ \varphi(y)}\ .$$

La variance est :

$$\sigma^2 = \frac{[y^r]\ \frac{\partial^2 \phi}{\partial x^2}\ (1,y)}{[y^r]\ \varphi(y)} + E - E^2$$

Démonstration

1) Soit $P_R(x_1, \ldots, x_d) = \prod_{\ell=1}^{d} \left(\frac{1 + p_\ell . x_\ell}{1 + p_\ell}\right)$, où p_ℓ est la probabilité de la ligne x_ℓ :

$$P_R = \sum_R P(R).R\ .$$

Pour obtenir $P[|s(R)| = t \;/\; |R| = r]$, nous remplaçons, dans le polynôme formel $x_{i_1} \ldots x_{i_r}$, représentant R dans P_R, chaque x_i par

$x^{1(x_i \in \Delta)} . y$. Nous obtenons donc : $x^{|s(R)|} \times y^{|R|}$. Cette opération appliquée à $P_R(x_1, \ldots, x_d)$ nous donne :

$$\prod_{\ell=1}^{d} \left(\frac{1 + p_\ell \, x^{1(x_\ell \in \Delta)} \, y}{1 + p_\ell} \right) = \sum_R P(R) \cdot x^{|s(R)|} \cdot y^{|R|}$$

$$= \sum_{t,r} P[|R| = r \text{ et } |s(R)| = t] \cdot x^t y^r \quad .$$

Soit $\psi(x,y) = \prod_{\ell=1}^{d} \left(\frac{1 + p_\ell \, x^{1(x_\ell \in \Delta)} \, y}{1 + p_\ell} \right)$:

$$P[|R| = r \text{ et } |s(R)| = t] = [x^t y^r] \, \psi(x,y) \; .$$

$$\text{Donc : } P[|s(R)| = t \;/\; |R| = r] = \frac{[x^t y^r] \, \psi(x,y)}{P[|R| = r]} \; .$$

$$\text{Or : } P[|R| = r] = [x^r] \, P_R(1, \ldots, 1) = [x^r] \, \psi(1,x)$$

$$\text{donc : } P[|s(R)| = t / |R| = r] = \frac{[x^t y^r] \, \psi(x,y)}{[y^r] \, \psi(1,y)} \; .$$

Soit, en posant $\varphi(y) = \prod_{\ell=1}^{d} (1 + p_\ell \, y)$

$$\text{et } \phi(x,y) = \prod_{\ell=1}^{d} (1 + p_\ell \cdot x^{1(x_\ell \in \Delta)} \, y) :$$

$$P[|s(R)| = t \;/\; |R| = r] = \frac{[x^t y^r] \, \phi(x,y)}{[y^r] \, \varphi(y)} \; .$$

2) Taille moyenne

$$E[|s(R)| \,/\, |R| = r] = \frac{[y^r]\ \frac{\partial \phi}{\partial x}(1,y)}{[y^r]\ \varphi(y)}$$

$$\cdot\ \frac{\partial \phi}{\partial x}(x,y) = \phi(x,y) \times \sum_{\ell=1}^{d} \frac{1_{(x_\ell \in \Delta)} \cdot p_\ell\ y}{1+p_\ell\ y\ x^{1_{(x_\ell \in \Delta)}}}$$

$$\Rightarrow \frac{\partial \phi}{\partial x}(1,y) = \varphi(y) \cdot \sum_{\ell=1}^{d} \frac{1_{(x_\ell \in \Delta)} \cdot p_\ell\ y}{1+p_\ell\ y} \quad .$$

Donc :

$$E[|s(R)| \,/\, |R| = r] = \frac{[y^{r-1}]\left\{\varphi(y) \cdot \sum_{\ell : x_\ell \in \Delta} \frac{p_\ell}{1+p_\ell\ y}\right\}}{[y^r]\quad (y)}$$

$$= \sum_{\ell=1}^{d} 1_{(x_\ell \in \Delta)} \cdot p_\ell - \frac{[y^r]\left\{\varphi(y) \cdot \sum_{\ell : x_\ell \in \Delta} \frac{1}{1+p_\ell\ y}\right\}}{y^r \quad (y)} \quad .$$

Soit : $P[\Delta] = \sum_{\ell=1}^{d} 1_{(x_\ell \in \Delta)} \cdot p_\ell$

$$= \sum_{\ell=1}^{d} 1_{x_\ell \in \Delta)} \cdot p(x_\ell) \ :$$

$$E[|s(R)| \,/\, |R| = r] = P[\Delta] - \frac{[y^r]\left\{\varphi(y) \cdot \sum_{\ell : x_\ell \in \Delta} \frac{1}{1+p_\ell\ y}\right\}}{[y^r]\ \varphi(y)} \quad .$$

3) Variance

$$\sigma^2 = \sum_{t=0}^{r} t(t-1)\, P[|s(R)| = t \;/\; |R| = r]$$

$$+ E[|s(R)|]\,(1 - E[|s(R)|])$$

$$\cdot \sum_{t=0}^{t} t(t-1)\, P[|s(R)| = t \;/\; |R| = r] = \frac{[y^r]\,\frac{\partial^2 \phi}{\partial x^2}(1,y)}{[y^r]\,\varphi(y)} \quad .$$

Donc :

$$\sigma^2 = \frac{[y^r]\,\frac{\partial^2 \phi}{\partial x^2}(1,y)}{[y^r]\,\varphi(y)} + E(|s(R)|) - E(|s(R)|)^2 \quad . \quad \blacksquare$$

C - RELATION AVEC DÉPENDANCE FONCTIONNELLE

Nous supposons ici que R est de la forme $R[A,B]$, avec $A \to B$.

Nous étudions d'abord la loi de taille de $s(R)$ dans le cas où l'ensemble Δ n'a aucune propriété particulière. Puis nous regardons, pour des lois uniformes ou non, les cas où Δ est de la forme :

$$\Delta_A \times \Delta_B \text{ , avec } \Delta_A \subseteq D_A$$
$$\Delta_B \subseteq D_B \text{ .}$$

I - LOI DE $|s(R)| / |R|$.

Notations

Pour tout a de D_A, nous définissons $\underline{\Delta_B(a)} = \{b \in D_B$, tel que $(a,b) \in \Delta\}$,

de probabilité $P[\Delta_B(a)] = f(a)$.

La probabilité que $s(R)$ soit de taille t, sachant que R est de taille r, est :

$$P[|s(R)| = t \ / \ |R| = r] = \frac{[x^t y^r]\ \phi(x,y)}{[y^r]\ \varphi(y)}$$

où $\varphi(y) = \prod_{a=1}^{d_A} (1 + p_a\, y)$

$$\phi(x,y) = \prod_{a=1}^{d_A} (1 + p_a\, y\, [x.f(a) + 1 - f(a)]) \text{ .}$$

La taille moyenne de $s(R)$, sachant que $|R| = r$, est :

$$E = \frac{[y^{r-1}] \left\{ \varphi(y) \sum_{a=1}^{d_A} \frac{p_a . f(a)}{1 + p_a\, y} \right\}}{[y^r]\ \varphi(y)} \quad .$$

La variance est :

$$\sigma^2 = \frac{[y^r]\ \frac{\partial^2 \phi}{\partial x^2}\ (1,y)}{[y^r]\ \varphi(y)} + E - E^2 \quad .$$

Démonstration

1° En identifiant la valeur de R : $\begin{bmatrix} x_{a_1} & y_{b_1} \\ \vdots & \vdots \\ x_{a_2} & y_{b_2} \end{bmatrix}$ et le polynôme formel

$x_{a_1}\, y_{b_1} \cdots x_{a_2}\, y_{b_2}$, nous avons pour fonction génératrice des valeurs de R :

$$f_R(x_1,\ldots,x_{d_A},\ y_1,\ldots,y_{d_B}) = \sum_R P(R).R$$

$$= \prod_{a=1}^{d_A} \left(\frac{1 + p_a\, x_a \left[\sum_{b=1}^{d_B} q_b\, y_b \right]}{1 + p_a} \right)$$

où p_a = Probabilité de x_a

et q_b = Probabilité de y_b .

Pour obtenir les valeurs de R , pour lesquelles $|s(R)|$ vaut t , et $|R|$ vaut r , nous remarquons qu'il suffit de remplacer, dans l'expression de R , $x_a\, y_b$ par $x^{1(x_a\, y_b \in \Delta)} . y$. Cette transformation appliquée à f_R nous donne, d'une part :

$$\sum_{R} P(R) \cdot x^{|s(R)|} \cdot y^{|R|} = \sum_{t,r} P[R, |R| = r \text{ et } |s(R)| = t]\, x^t y^r \quad ;$$

d'autre part :

$$\prod_{a=1}^{d_A} \left(\frac{1 + p_a \cdot \left[\sum_{b=1}^{d_B} q_b \, x^{1_{(x_a y_b \in \Delta)}} \right] y}{1 + p_a} \right)$$

Or : si $y_b \in \Delta_B(a)$: $1_{(x_a y_b \in \Delta)} = 1$

si $y_b \notin \Delta_B(a)$: $1_{(x_a y_b \in \Delta)} = 0$.

Donc :

$$\sum_{b=1}^{d_B} q_b \, x^{1_{(x_a y_b \in \Delta)}} = x \cdot \sum_{b=1}^{d_B} q_b \cdot 1_{(y_b \in \Delta_B(a))} + 1 - \sum_{b=1}^{d_B} q_b \cdot 1_{(y_b \in \Delta_B(a))}$$

$$= x.f(a) + 1 - f(a) \ .$$

Nous obtenons donc :

$$\sum_{t,r} x^t y^r \times P[|R| = r \text{ et } |s(R)| = t] = \prod_{a=1}^{d_A} \left(\frac{1 + p_a \, y[f(a).x + 1 - f(a)]}{1 + p_a} \right) \ .$$

Par ailleurs :

$$P[|s(R)| = t \ / \ |R| = r] = \frac{P[|s(R)| = t \text{ et } |R| = r]}{P[|R| = r]}$$

et :

$$P[|R| = r] = [y^r] \left\{ \prod_{a=1}^{d_A} \left(\frac{1 + p_a \, y}{1 + p_a} \right) \right\} \ .$$

D'où :

$$P[|s(R)| = t \,/\, |R| = r] = \frac{[x^t y^r] \left\{ \prod_{a=1}^{d_A} (1 + p_a\, y[f(a).x + 1 - f(a)]) \right\}}{[y^r] \left\{ \prod_{a=1}^{d_A} (1 + p_a\, y) \right\}} \quad .$$

2) Taille moyenne

$$E[|s(R)| \,/\, |R| = r] = \sum_{t=0}^{r} t.P[|s(R)| = t \,/\, |R| = r]$$

$$= \sum_{t=0}^{r} t. \frac{[x^t y^r]\, \phi(x,y)}{[y^r]\, \varphi(y)}$$

$$= \frac{[y^r] \frac{\partial \phi}{\partial x}(1,y)}{[y^r]\, \varphi(y)}$$

$$\phi(x,y) = \prod_{a=1}^{d_A} (1 + p_a\, y[x\, f(a) + 1 - f(a)])$$

$$\frac{\partial \phi}{\partial x}(x,y) = \phi(x,y) \times \sum_{a=1}^{d_A} \frac{p_a\, y\, f(a)}{1 + p_a\, y[x\, f(a) + 1 - f(a)]}$$

$$\Rightarrow \frac{\partial \phi}{\partial x}(1,y) = y \times \varphi(y) \times \sum_{a=1}^{d_A} \frac{p_a . f(a)}{1 + p_a\, y} \quad .$$

Donc : $E[|s(R)| \,/\, |R| = r] = \dfrac{[y^{r-1}] \left\{ \varphi(y) \sum_{a=1}^{d_A} \frac{p_a \,.\, f(a)}{1 + p_a\, y} \right\}}{[y^r]\, \varphi(y)} \quad .$

3) Variance

$$E(|s(R)|^2 \;/\; |R| = r) = \sum_{t=0}^{r} t(t-1)\,\frac{[x^t y^r]\phi(x,y)}{[y^r]\,\varphi(y)} + \sum_{t=0}^{r} t.\,\frac{[x^t y^r]\phi(x,y)}{[y^r]\,\varphi(y)}$$

$$= \frac{[y^r]\,\frac{\partial \phi}{\partial x}(1,y)}{[y^r]\,\varphi(y)} + E[|s(R)| \;/\; |R| = r]$$

et $$\sigma^2 = E(|s(R)|^2 \;/\; |R| = r) - E(|s(R)| \;/\; |R| = r)^2$$

d'où le résultat. ∎

II - LOIS UNIFORMES

1) Soit $\Delta = \Delta_A \times \Delta_B$, $\delta_A = |\Delta_A|$, $\delta_B = |\Delta_B|$,

$d_A = |D_A|$ et $d_B = |D_B|$

La probabilité que $s(R)$ soit de taille t , sachant que R est de taille r , est :

$$P[|s(R)| = t \;/\; |R| = r] = \frac{[x^t y^r]\phi(x,y)}{[y^r]\,\varphi(y)}$$

où $\varphi(y) = (1 + d_B\, y)^{d_A}$

$$\phi(x,y) = (1 + d_B\, y)^{d_A - \delta_A}\,(1 + (d_B - \delta_B)\, y + \delta_B\, xy)^{\delta_A}\ .$$

La taille moyenne de $s(R)$ est :

$$r.\,\frac{\delta_A \cdot \delta_B}{d_A \cdot d_B}$$

et la variance :

$$\sigma^2 = r \cdot \frac{\delta_A}{d_A} \cdot \frac{\delta_B}{d_B} \left[1 - r \cdot \frac{\delta_A}{d_A} \cdot \frac{\delta_B}{d_B} + (r-1) \frac{\delta_A - 1}{d_A - 1} \cdot \frac{\delta_B}{d_B} \right] .$$

Démonstration

En prenant comme fonction génératrice des valeurs de

$$R : \frac{1}{2^{d_A d_B}} \cdot \prod_{a=1}^{d_A} (1 + x_a \cdot \sum_{b=1}^{d_B} y_b) ,$$

un raisonnement analogue à celui fait dans le cas général donne :

$$P[|s(R)| = t \ / \ |R| = r] = \frac{[x^t y^r] \phi(x,y)}{[y^r] \varphi(y)}$$

où $\phi(x,y) = (1 + d_B y)^{d_A - \delta_A} (1 + (d_B - \delta_B) y + \delta_B xy)^{\delta_A}$

$$\varphi(y) = \phi(1,y) = (1 + d_B y)^{d_A} .$$

Ceci nous permet d'obtenir les moments de $|s(R)|$: taille moyenne E, et variance σ^2 .

$$E = \frac{[y^r] \frac{\partial \phi}{\partial x} (1,y)}{[y^r] \varphi(y)}$$

$$\sigma^2 = \frac{[y^r] \frac{\partial^2 \phi}{\partial x^2} (1,y)}{[y^r] \varphi(y)} + E - E^2$$

$$\cdot \frac{\partial \phi}{\partial x} (x,y) = (1 + d_B y)^{d_A - \delta_A} \cdot \delta_A \delta_B y \cdot (1 + (d_B - \delta_B) y + \delta_B x y)^{\delta_A - 1}$$

$$\Rightarrow \frac{\partial \phi}{\partial x}(1,y) = \delta_A \delta_B \, y \cdot (1 + d_B \, y)^{d_A - 1}$$

$$\text{donc : } E = \frac{\delta_A \cdot \delta_B \cdot {d_B}^{r-1} \binom{d_A - 1}{r-1}}{{d_B}^r \binom{d_A}{r}} = \frac{r \cdot \delta_A \cdot \delta_B}{d_A \, d_B}$$

$$\cdot \frac{\partial^2 \phi}{\partial x^2}(x,y) = (1 + d_B \, y)^{d_A - \delta_A} \cdot \delta_A \cdot (\delta_A - 1) \cdot \delta_B^2 \, y^2 (1 + (d_B - \delta_B) \, y + \delta_B \, xy)^{\delta_A - 2}$$

$$\Rightarrow \frac{\partial^2 \phi}{\partial x^2}(1,y) = (1 + d_B \, y)^{d_A - 2} \cdot \delta_A \cdot (\delta_A - 1) \cdot {\delta_B}^2 \cdot y^2 \quad .$$

$$\text{D'où : } \frac{[y^r] \frac{\partial^2 \phi}{\partial x^2}(1,y)}{[y^r] \, \varphi(y)} = \frac{\delta_A \cdot (\delta_A - 1) \cdot {\delta_B}^2 \cdot {d_B}^{r-2} \cdot \binom{d_A - 2}{r-2}}{{d_B}^r \cdot \binom{d_A}{r}}$$

$$= r(r-1) \cdot \frac{\delta_A}{d_A} \cdot \frac{\delta_A - 1}{d_A - 1} \cdot \left(\frac{\delta_B}{d_B}\right)^2$$

$$\text{et } \sigma^2 = r \cdot \frac{\delta_A}{d_A} \cdot \frac{\delta_B}{d_B} \left[(r-1) \cdot \frac{\delta_A - 1}{d_A - 1} \cdot \frac{\delta_B}{d_B} + 1 - r \cdot \frac{\delta_A}{d_A} \cdot \frac{\delta_B}{d_B} \right] \quad .$$

2) Soit $\underline{\Delta = \Delta_A \times D_B}$

Ceci correspond au cas où aucune condition n'est imposée à l'attribut B. Nous avons alors :

$$P[|s(R)| = t \;/\; |R| = r] = \frac{\binom{\delta_A}{t} \binom{d_A - \delta_A}{r - t}}{\binom{d_A}{r}}$$

$$E[|s(R)| \;/\; |R| = r] = r \cdot \frac{\delta_A}{d_A}$$

$$\sigma^2 = \frac{r\,\delta_A(d_A - \delta_A)\,(d_A - r)}{d_A^{\,2}(d_A - 1)} \quad .$$

Démonstration

Les résultats précédents, appliqués au cas $D_B = \Delta_B$, donnent la loi de probabilité et ses moments. Une autre manière de calculer ici la loi de $|s(R)|$ est de remarquer que, aucune condition n'étant imposée à l'attribut B , et l'attribut A étant la clé de R , tout se passe comme si nous avions une relation à un seul attribut A , donc sans dépendance fonctionnelle, et d'appliquer les résultats obtenus dans ce cas. ■

3) Soit $\Delta = D_A \times \Delta_B$

(Cas où aucune condition n'est imposée à l'attribut A , clé de la relation).

$$P[|s(R)| = t \;/\; |R| = r] = \frac{\binom{d_A}{t}\binom{d_A - t}{r - t}}{\binom{d_A}{r}} \left(\frac{\delta_B}{d_B}\right) \cdot \left(1 - \frac{\delta_B}{d_B}\right)^{r-t}$$

$$E[|s(R)| \;/\; |R| = r] = r.\frac{\delta_B}{d_B}$$

$$\sigma^2[|s(R)| \;/\; |R| = r] = r.\frac{\delta_B}{d_B}\left(1 - \frac{\delta_B}{d_B}\right)$$

Démonstration

Il suffit d'appliquer les résultats de 1) au cas où $D_A = \Delta_A$. ■

III - LOIS GENERALES

1) Soit $\Delta = \Delta_A \times \Delta_B$, $\Delta_B \subseteq D_A$ et $\Delta_B \subseteq D_B$.

La probabilité que $s(R)$ soit de taille t , sachant que R est de taille r , est :

$$P[|s(R)| = t \;/\; |R| = r] = \frac{[x^t y^r]\ \phi(x,y)}{[y^r]\ \varphi(y)}$$

où $$\varphi(y) = \prod_{a=1}^{d_A} (1 + p_a\, y)$$

$$\phi(x,y) = \varphi(y) \times \prod_{a \in \Delta_a} \left(\frac{1 + p_a\, y[x.P(\Delta_B) + 1 - P(\Delta_B)]}{1 + p_a\, y} \right) .$$

La taille moyenne de $s(R)$ est :

$$E = \frac{[y^{r-1}] \left\{ \varphi(y) . \sum_{a \in \Delta_a} \frac{p_a}{1 + p_a\, y} \right\} \times P(\Delta_B)}{[y^r]\ \varphi(y)}$$

et la variance :

$$\sigma^2 = \frac{[y^r]\ \frac{\partial^2 \phi}{\partial x^2}\ (1,y)}{[y^r]\ (y)} + E - E^2 .$$

Démonstration

Il suffit d'appliquer les résultats obtenus dans le cas de Δ quelconque : la seule différence est qu'ici : pour tout a , $\Delta_B(a) = \Delta_B$. ▪

2) Soit $\Delta = \Delta_A \times D_B$

(Cas où aucune condition n'est imposée à l'attribut B).

$$P[|s(R)| = t \;/\; |R| = r] = \frac{[x^t y^r]\left\{\prod_{a \notin \Delta_a} (\ell + p_a\, y).\prod_{a \in \Delta_a} (1 + p_a\, xy)\right\}}{[y^r]\left\{\prod_{a \in D_A} (1 + p_a\, y)\right\}}$$

$$E = \frac{[y^{r-1}]\left\{\varphi(y).\sum_{a \in \Delta_a} \frac{p_a}{1 + p_a\, y}\right\}}{[y^r]\,\varphi(y)}$$

3) Soit $\Delta = D_A \times \Delta_B$

(Cas où aucune condition n'est imposée à l'attribut A , clé de la relation).

$$P[|s(R)| = t \;/\; |R| = r] = \frac{[x^t y^r]\left\{\prod_{a \in D_A} (1 + p_a\, y(x\, P[\Delta_B] + 1 - P[\Delta_B])\right\}}{[y^r]\left\{\prod_{a \in D_A} (1 + p_a\, y)\right\}}$$

$$E[|s(R)| \;/\; |R| = r] = \frac{[y^{r-1}]\left\{\varphi(y)\sum_{a \in D_A} \frac{p_a}{1 + p_a\, y}\right\} P(\Delta_B)}{[y^r]\,\varphi(y)} \quad .$$

CHAPITRE III

UNION, INTERSECTION, DIFFERENCE

Nous étudions dans ce chapitre les opérations d'union, intersection et différence portant sur deux valeurs R et S d'un même schéma de relation. Ces valeurs sont supposées indépendantes, ce qui entraîne en particulier que la probabilité du couple (R, S) est égale au produit des probabilités de R et de S.

Nous supposerons que le schéma de relation commun à R et S ne comporte pas de dépendance fonctionnelle, ou bien comporte une dépendance fonctionnelle "simple".

Nous étudierons en détail la loi de taille de l'intersection de R et S, en supposant connues les tailles de R et S. Les tailles des autres opérations s'en déduisent comme suit :

$$|R \cup S| = |R| + |S| - |R \cap S|$$

$$|R \setminus S| = |R| - |R \cap S|$$

$$|S \setminus R| = |S| - |R \cap S|$$

$$|R \triangle S| = |R| + |S| - 2|R \cap S|$$

Nous obtenons, pour des lois uniformes sur les domaines des attributs des relations, la loi de taille de l'intersection, sa taille moyenne et sa variance.

Nous étudions aussi le cas de lois de probabilités plus générales, en supposant que $P(R)$ est proportionnel à $\prod_{\ell \in R} P(\ell)$.

Nous donnons la loi de taille de l'intersection et sa taille moyenne.

A - RELATIONS SANS DEPENDANCE FONCTIONNELLE

I - LOIS UNIFORMES

Loi de $|R \cap S| / |R|$ et $|S|$:

La probabilité que la taille de $R \cap S$ soit t, sachant que la taille de R est r, et celle de S, s, est :

$$P[|R \cap S| = t / |R| = r \text{ et } |S| = s] = \frac{\binom{d-r}{s-t}\binom{r}{t}}{\binom{d}{s}}$$

La taille moyenne de $R \cap S$ est alors $\frac{r\,s}{d}$

et la variance : $\sigma^2 = \frac{rs(d-r)(d-s)}{d^2(d-1)}$

si r et $s \ll d$: $\sigma^2 \simeq \frac{rs}{d}\left(1 - \frac{r+s-1}{d}\right)$.

Démonstration

Nous indiquerons deux manières de trouver la loi de $|R \cap S|$: par le dénombrement ou par les fonctions génératrices.

1°) Dénombrement

La loi sur l'ensemble des lignes pouvant être prises par R ou S étant uniforme, la loi sur l'ensemble des valeurs de R ou S, de taille donnée, l'est aussi.

Donc la probabilité que $R \cap S$ soit de taille t, pour R de taille r et S de taille s, est le rapport du nombre de cas favorables au nombre total de cas possibles :

$$P[|R \cap S| = t \;/\; |R| = r \quad \text{et} \quad |S| = s] = \frac{N_t}{N}$$

où . N est le nombre total de couples (R,S), vérifiant $|R| = r$ et $|S| = s$. Les valeurs prises par R et S étant indépendantes, N = Nombre de valeurs de R × Nombre de valeurs de S

$$N = \binom{d}{r} \cdot \binom{d}{s}$$

. N_t est le nombre de couples (R,S) tels que $|R| = r$, $|S| = s$, et $|R \cap S| = t$. Pour calculer N_t, nous remarquons que nous pouvons construire un tel couple (R,S) comme suit :

- d'abord choisir R, de taille r
- puis fixer les t éléments de R, qui seront aussi dans S
- enfin compléter S par $(s-t)$ éléments non déjà dans R.

Donc :

$$N_t = \binom{d}{r} \cdot \binom{r}{t} \cdot \binom{d-r}{s-t} \quad .$$

Nous avons donc :

$$P[|R \cap S| = t \;/\; |R| = r \quad \text{et} \quad |S| = s] = \frac{\binom{r}{t}\binom{d-r}{s-t}}{\binom{d}{s}} \quad .$$

Remarque :

$$\frac{\binom{r}{t}\binom{d-r}{s-t}}{\binom{d}{s}} = \frac{(d-r)!\ (d-s)!\ r!\ s!}{d!\ t!\ (r-t)!\ (s-t)!\ (d+t-r-s)!}$$

$$= \frac{\binom{s}{t}\binom{d-s}{r-t}}{\binom{d}{r}}$$

r et s jouent bien des rôles identiques dans l'expression de $P[|R \cap S| = t \,/\, |R| = r \text{ et } |S| = s]$.

2°) Fonction génératrice

Nous identifierons la valeur de $R : \{x_{i_1},\ldots,x_{i_r}\}$ et le polynôme formel $x_{i_1} \ldots x_{i_r}$.

Nous pouvons donc représenter l'ensemble des valeurs de R par :

$$\sum_{r \geqslant 0} \sum_{i_1 \ldots i_r} x_{i_1} \ldots x_{i_r} = \prod_{i=1}^{d} (1 + x_i)$$

Soit $P(x_1,\ldots,x_d)$ ce polynôme.

Nous cherchons d'abord, à S fixé, le nombre de valeurs de R, de taille r, telles que $|R \cap S| = t$. Pour cela, nous remplaçons dans le monôme représentant $R : \prod_{i=1}^{d} x_i^{1_{(x_i \in R)}}$, chaque x_i par $x^{1_{(x_i \in S)}}.y$. Nous obtenons :

$$x^{\sum_{i=1}^{d} 1_{(x_i \in R)}\, 1_{(x_i \in S)}}.y^{\sum_{i=1}^{d} 1_{(x_i \in R)}} = x^{|R \cap S|}\, y^{|R|} .$$

Soit $\varphi(x,y)$ la fonction obtenue par le remplacement de x_i par $x^{1(x_i \in S)}.y$ dans P :

$$\varphi(x,y) = \prod_{i=1}^{d} (1 + x^{1(x_i \in S)}.y)$$

$$= (1 + xy)^s \cdot (1 + y)^{d-s} .$$

Par ailleurs, $\varphi(x,y) = \sum_{r,t} N_{r,t}\, x^t\, y^r$ où $N_{r,t}$ est le nombre de valeurs de R , telles que $|R| = r$ et $|R \cap S| = t$.

Si N est le nombre de valeurs de R , telles que

$$|R| = r : P[|R \cap S| = t / |R| = r \text{ et } S] = \frac{N_{r,t}}{N}$$

$$P[|R \cap S| = t / |R| = r \text{ et } S] = \frac{1}{\binom{d}{r}} [x^t\, y^r]\, \varphi(x,y)$$

$$= \frac{1}{\binom{d}{r}} \cdot \binom{s}{t} \binom{d-s}{r-t}$$

Pour obtenir $P[|R \cap S| = t / |R| = r \text{ et } |S| = s]$, nous remarquons que : $P[|R \cap S| = t / |R| = r \text{ et } |S| = s] =$

$$= \sum_{\substack{S_0 \text{ , tel que} \\ |S_0| = s}} P[|R \cap S| = t / |R| = r \text{ et } S = S_0].P[S = S_0 / |S| = s]$$

Or $P[S = S_0 / |S| = s] = \dfrac{1}{\text{Nombre de } S, \text{ tel que } |S| = s}$

et $P[|R \cap S| = t / |R| = r \text{ et } |S| = S_0]$ ne dépend pas ici de S_0 .

Donc :

$$P[|R \cap S| = t / |R| = r \text{ et } |S| = s] = \frac{\binom{s}{t}\binom{d-s}{r-t}}{\binom{d}{r}}$$

$$= [x^t y^r] \ \phi(x,y)$$

où $\phi(x,y) = \frac{1}{\binom{d}{r}} \varphi(x,y) = \frac{1}{\binom{d}{r}} (1+xy)^s (1+y)^{d-s}$

3°) Taille moyenne et variance

Nous pouvons obtenir les moments de la loi de $|R \cap S|$ directement, en remplaçant $P[|R \cap S| = t / |R| = r \text{ et } |S| = s]$ par sa valeur explicite. Nous donnerons ici une autre démonstration, utilisant la fonction génératrice de $|R \cap S|$:

$$E[|R \cap S| / |R| = r \text{ et } |S| = s] = \sum_{t \geq 0} P[|R \cap S| = t / |R| = r \text{ et } |S| = s].t$$

Or :

$$\frac{\partial \phi}{\partial X}(X,Y) = \sum_{r,t \geq 0} t.P[|R \cap S| = t / |R| = r \text{ et } |S| = s].x^{t-1} y^r \ ,$$

et

$$\frac{\partial \phi}{\partial X}(1,Y) = \sum_r [\sum_t t.P[|R \cap S| = t / |R| = r \text{ et } |S| = s].y^r] \ .$$

Donc $E[|R \cap S| / |R| = r \text{ et } |S| = s] = [y^r] \frac{\partial \phi}{\partial X}(1,Y)$.

$$\frac{\partial \phi}{\partial X}(X,Y) = \frac{s.y}{\binom{d}{r}} (1+y)^{d-s} (1+xy)^{s-1}$$

$$\Rightarrow \quad \frac{\partial \phi}{\partial X}(1,Y) = \frac{sy}{\binom{d}{r}} (1+y)^{d-1}$$

$$= \frac{sy}{\binom{d}{r}} \sum_{u=0}^{d-1} \binom{d-1}{u} y^u \quad .$$

Donc $$[y^r] \frac{\partial \phi}{\partial X}(1,y) = \frac{s}{\binom{d}{r}} \times \binom{d-1}{r-1} = \frac{rs}{d} \quad .$$

Notons σ^2 la variance de $|R \cap S|$, à $|R| = r$ et $|S| = s$:

$$\sigma^2 = \sum_t t^2 \, P[|R \cap S| = t \,/\, |R| = \text{ et } |S| = s] - (E[|R \cap S| \,/\, |R| = r \text{ et } |S| = s])^2$$

$$\sigma^2 = \sum_t t(t-1) \, P[|R \cap S| = t \,/\, |R| = r \text{ et } |S| = s] + E[|R \cap S| \,/\, |R| = r \text{ et } |S| = s]$$

$$- (E[|R \cap S| \,/\, |R| = r \text{ et } |S| = s])^2 \quad .$$

Nous avons, comme ci-dessus :

$$\sum_t t(t-1) \, P[|R \cap S| = t \,/\, |R| = r \text{ et } |S| = s] = [y^r] \frac{\partial^2 \phi}{\partial X^2}(1,Y)$$

$$\frac{\partial^2 \phi}{\partial X^2}(X,Y) = \frac{s(s-1)y^2}{\binom{d}{r}} (1+y)^{d-s} (1+xy)^{s-2} \quad .$$

Donc $$\frac{\partial^2 \phi}{\partial X^2}(1,Y) = \frac{s(s-1)\, y^2}{\binom{d}{r}} (1+y)^{d-2}$$

et $$[y^r] \frac{\partial^2 \phi}{\partial X^2}(1,Y) = \frac{s(s-1)\binom{d-2}{r-2}}{\binom{d}{r}} = \frac{rs(r-1)(s-1)}{d(d-1)}$$

Nous en tirons :

$$\sigma^2 = \frac{rs(r-1)(s-1)}{d(d-1)} + \frac{rs}{d} - \left(\frac{rs}{d}\right)^2$$

$$\sigma^2 = \frac{rs}{d} \times \frac{d^2 - (r+s)d + rs}{d(d-1)}$$

$$\sigma^2 = \frac{rs(d-r)(d-s)}{d^2(d-1)}$$

Si r et s sont petits devant d,

$$\sigma^2 = \frac{rs}{d}\left(1-\frac{r}{d}\right)\left(1-\frac{s}{d}\right)\left(1-\frac{1}{d}\right)^{-1}$$

$$\sigma^2 \simeq \frac{rs}{d}\left(1-\frac{r}{d}-\frac{s}{d}+\frac{1}{d}\right) = \frac{rs}{d}\left(1-\frac{r+s-1}{d}\right).$$

II - LOIS GENERALES

1) Loi de $|R \cap S| / (|R| = r \text{ et } S = S_0)$

La probabilité que la taille de $R \cap S$ soit t, sachant que la taille de R est r, et que S prend la valeur S_0, est :

$$P[|R \cap S| = t / |R| = r \text{ et } S = S_0]$$

$$= \frac{[x^t y^r] \prod_{i=1}^{d} (1 + p_i x^{1_{(x_i \in S_0)}} . y)}{[x^r] \prod_{i=1}^{d} (1 + p_i x)} .$$

La taille moyenne de $R \cap S$ est :

$$|S_0| - \sum_{u=0}^{r} \sum_{i=1}^{d} (- P_i)^u . 1_{(x_i \in S_0)} \frac{[x^{r-u}] \prod_{i=1}^{d} (1 + p_i x)}{[x^r] \prod_{i=1}^{d} (1 + p_i x)}$$

Démonstration

1°) A chaque valeur $\{x_{i_1} \ldots x_{i_r}\}$ prise par R, nous associons $p_{i_1} \ldots p_{i_r} \cdot x_{i_1} \ldots x_{i_r}$, où p_{i_k} est la probabilité de la ligne x_{i_k}. L'ensemble des valeurs prises par R est donc représenté par :

$$P(x_1, \ldots, x_d) = \sum_{r \geq 0} \sum_{i_1 \ldots i_r} p_{i_1} \ldots p_{i_r} \cdot x_{i_1} \ldots x_{i_r}$$

$$= \prod_{i=1}^{d} (1 + p_i x_i) .$$

Pour obtenir la fonction génératrice des valeurs de R : $\sum_R P(R).R$, il suffit de prendre :

$$\frac{P(x_1,\ldots,x_d)}{P(1,\ldots,1)} = f(x_1,\ldots,x_d) :$$

$$f(x_1,\ldots,x_d) = \sum_R P(R).R = \prod_{i=1}^{d} \left(\frac{1+p_i\, x_i}{1+p_i}\right)$$

$$* \quad |R \cap S| = \sum_{i=1}^{d} 1_{(x_i \in R \cap S)} = \sum_{i=1}^{d} 1_{(x_i \in R)}\, 1_{(x_i \in S)} .$$

La probabilité que $|R \cap S|$ soit de taille t , à $S = S_0$ fixé, s'obtient donc en remplaçant, dans $f(x_1,\ldots,x_d)$, chaque x_i par $x^{1(x_i \in S_0)}.y$

Soit $\varphi_{S_0}(x,y)$ la fonction ainsi obtenue :

$$\varphi_{S_0}(x,y) = \sum_{R_0} P(R_0) \,.\, x^{|R_0 \cap S_0|} \,.\, y^{|R_0|}$$

$$= \sum_{r,t} \Big[\sum_{\substack{R_0 \text{ , tel que} \\ |R_0| = r \text{ et} \\ |R_0 \cap S_0| = t}} P(R_0) \Big]\, x^t\, y^r$$

$$= \sum_{r,t} P[|R \cap S_0| = t \text{ et } |R| = r \,/\, S = S_0]\, x^t\, y^r .$$

Donc : $P[|R \cap S_0| = t \text{ et } |R| = r \,/\, S = S_0] = [x^t y^r]\, \varphi_{S_0}(x,y)$

et $$P[|R \cap S_0| = t \,/\, |R| = r \text{ et } S = S_0] = \frac{[x^t\, y^r]\, \varphi_{S_0}(x,y)}{P[|R| = r \,/\, S = S_0]} .$$

Or, les valeurs prises par R et S étant indépendantes :

$$P[|R| = r/S = S_0] = P[|R| = r]$$

$$= [x^r]\ f(x,\ldots,x)$$

$$= [x^r] \prod_{i=1}^{d} \frac{(1+p_i\ x)}{1+P_i}\ .$$

D'autre part, $\varphi_{S_0}(x,y) = \prod_{i=1}^{d} \frac{1+p_i\ .\ x^{1_{(x_i \in S_0)}}\ .\ y}{1+p_i}$.

Donc :

$$P[|R \cap S_0| = t\ /\ |R| = r \text{ et } S = S_0]$$

$$= \frac{[x^t\ y^r] \prod_{i=1}^{d} (1+p_i\ x^{1_{(x_i \in S_0)}}\ .\ y)}{[x^r] \prod_{i=1}^{d} (1+p_i\ x)}\ .$$

2°) Taille moyenne de $R \cap S_0$:

$$E[|R \cap S|\ /|R| = r \text{ et } S = S_0] = \sum_{t \geqslant 0} t.P[|R \cap S| = t/|R| = r \text{ et } S = S_0]$$

$$= \frac{\sum_{t \geqslant 0} t.[x^t\ y^r] \prod_{i=1}^{d} (1+p_i\ x^{1_{(x_i \in S_0)}}\ y)}{[x^r] \prod_{i=1}^{d} (1+p_i\ x)}\ .$$

Pour calculer E , nous remarquons que, en notant

$$\phi(x,y) = \prod_{i=1}^{d} (1+p_i\ x^{1_{(x_i \in S_0)}}\ y) :$$

$$\sum_{t \geqslant 0} t.[x^t\ y^r]\ \phi(x,y) = [y^r]\ \frac{\partial \phi}{\partial x}\ (1,y)\ .$$

Or :

$$\frac{\partial \phi}{\partial x}(x,y) = \phi(x,y) \times \sum_{i=1}^{d} \frac{1_{(x_i \in S_0)} \cdot p_i\, y}{1 + p_i\, x^{1_{(x_i \in S_0)}} y}$$

et

$$\frac{\partial \phi}{\partial x}(1,y) = \prod_{j=1}^{d} (1 + p_j\, y) \times \sum_{i=1}^{d} 1_{(x_i \in S_0)} \left(1 - \frac{1}{p_i\, y+1}\right)$$

$$= |S_0| \prod_{j=1}^{d} (1 + p_j\, y) - \sum_{i=1}^{d} 1_{(x_i \in S_0)} \prod_{j \neq i} (1 + p_j\, y)$$

donc

$$[y^r] \frac{\partial \phi}{\partial x}(1,y) = |S_0| \cdot [y^r] \prod_{j=1}^{d} (1 + p_j\, y)$$

$$- \sum_{i=1}^{d} 1_{(x_i \in S_0)} [y^r] \prod_{j \neq i} (1 + p_j\, y)$$

$$[y^r] \prod_{j \neq i} (1 + p_j\, y) = [y^r]\left(\frac{\phi(1,y)}{1+p_i\, y}\right) = \frac{1}{r!} \left(\frac{\phi(1,y)}{1+p_i\, y}\right)^{(r)} (0)$$

$$= \frac{1}{r!} \sum_{u=0}^{r} \binom{r}{u} \left(\frac{\partial^{(r-u)} \phi(1,y)}{\partial y^{(r-u)}}\right) \left(\frac{1}{1+p_i\, y}\right)^{(u)} (0)$$

$$= \frac{1}{r!} \sum_{u=0}^{r} \binom{r}{u} \cdot [(r-u)!\, [y^{r-u}] \prod_{j=1}^{d} (1 + p_j\, y)] \cdot (-p_i)^u\, u!$$

$$= \sum_{u=0}^{r} (-p_i)^u \times [y^{r-u}] \prod_{j=1}^{d} (1 + p_j\, y) \cdot$$

Nous en tirons, en reportant dans l'expression de la taille moyenne et en réordonnant :

$$E = |S_0| - \sum_{u=0}^{r} \frac{[y^{r-u}] \prod_{j=1}^{d} (1 + p_j\, y)}{[y^r] \prod_{j=1}^{d} (1 + p_j\, y)} \sum_{i=1}^{d} (-\, p_i)^u \,.\, 1_{(x_i \in S_0)} \quad . \quad \blacksquare$$

2) Loi de $|R \cap S|/|R| = r$ et $|S| = s$.

La probabilité que $R \cap S$ soit de taille t , sachant que R est de taille r et S de taille s , est :

$$P[|R \cap S| = t \,/\, |R| = r \text{ et } |S| = s]$$

$$= \frac{[x^t\, y^r\, z^s] \prod_{i=1}^{d} (1 + p_i(y+z) + p_i^{\,2}\, xyz)}{[y^r] \prod_{i=1}^{d} (1 + p_i\, y) \times [y^s] \prod_{i=1}^{d} (1 + p_i\, y)}$$

La taille moyenne de $R \cap S$ est :

$$\frac{\sum_{i=1}^{d} [x^r]\, (\varphi - \varphi_i).[x^s]\, (\varphi - \varphi_i)}{[x^r]\varphi \,.\, [x^s]\varphi}$$

où $\varphi(x) = \prod_{i=1}^{d} (1 + p_i\, x)$

et $\varphi_i(x) = \prod_{j \neq i} (1 + p_j\, x)$.

Démonstration

1°) $P[|R \cap S| = t/|R| = r \text{ et } |S| = s]$

$$= \sum_{S_0 \text{ , tel que } |S_0| = s} P[|R \cap S| = t/|R| = r \text{ et } S = S_0].\, P[S = S_0/|S| = s] \,.$$

. Nous avons calculé : $P[|R \cap S| = t/|R| = r \text{ et } S = S_0]$

$$\cdot \quad P[S = S_0 / |S| = s] = \frac{\prod_{i=1}^{d} p_i^{1(x_i \in S_0)}}{\sum_{i_1 \ldots i_s} p_{i_1} \ldots p_{i_s}} = \frac{\prod_{i=1}^{d} p_i^{1(x_i \in S_0)}}{[x^s] \prod_{i=1}^{d} (1 + p_i x)} \quad .$$

Donc $\quad P[|R \cap S| = t / |R| = r \text{ et } |S| = s]$

$$= \sum_{S_0, tq\ |S_0| = s} \frac{[x^t y^r] \prod_{i=1}^{d} (1 + p_i x^{1(x_i \in S_0)} y) \times \prod_{j=1}^{d} p_j^{1(x_j \in S_0)}}{[x^r] \prod_{i=1}^{d} (1 + p_i x) \times [x^s] \prod_{i=1}^{d} (1 + p_i x)}$$

$$\cdot \sum_{S_0, tq\ |S_0| = s} [x^t y^r] \prod_{i=1}^{d} (1 + p_i x^{1(x_i \in S_0)} y) \cdot \prod_{j=1}^{d} p_j^{1(x_j \in S_0)}$$

$$= \sum_{i_1 \ldots i_s} [x^t y^r] \left\{ p_{i_1} \ldots p_{i_s} \prod_{k=1}^{s} \frac{(1 + p_{i_k} x y)}{1 + p_{i_k} y} \times \prod_{\ell=1}^{d} (1 + p_\ell y) \right\}$$

$$= [x^t y^r] \left\{ \prod_{i=1}^{d} (1 + p_\ell y) \times \sum_{i_1 \ldots i_s} \prod_{k=1}^{s} p_{i_k} \frac{(1 + p_{i_k} x y)}{1 + p_{i_k} y} \right\} \quad .$$

Or : $\quad \sum_{i_1 \ldots i_s} \alpha_{i_1} \ldots \alpha_{i_s} = [z^s] \prod_{i=1}^{d} (1 + \alpha_i z) \quad .$

En remplaçant α_i par $\frac{p_i (1 + p_i x y)}{1 + p_i y}$, nous obtenons :

$$\sum_{S_0, tq\ |S_0| = s} [x^t y^r] \prod_{i=1}^{d} (1 + p_i x^{1(x_i \in S_0)} y) \prod_{j=1}^{d} p_j^{1(x_j \in S_0)}$$

$$= [x^t y^r] \left\{ \prod_{\ell=1}^{d} (1 + p_\ell y) \times [z^s] \prod_{i=1}^{d} \left(1 + \frac{(1 + p_i xy) p_i z}{1 + p_i y} \right) \right\}$$

$$= [x^t\, y^r\, z^s] \left\{ \prod_{i=1}^{d} (1 + p_i\, y + p_i\, z + p_i^2\, x\, y\, z) \right\}$$

$$= [x^t\, y^r\, z^s] \left\{ \prod_{i=1}^{d} (1 + p_i (y+z) + p_i^2\, x\, y\, z) \right\}$$

Donc :

$$P[|R \cap S| = t \,/\, |R| = r \text{ et } |S| = s]$$

$$= \frac{[x^t\, y^r\, z^s] \prod_{i=1}^{d} [1 + p_i (y+z) + p_i^2\, x\, y\, z]}{[y^r] \prod_{i=1}^{d} (1 + p_i\, y)\ [z^s] \prod_{i=1}^{d} (1 + p_i\, z)}$$

2°) Taille moyenne

Notons $f(x,y,z) = \prod_{i=1}^{d} (1 + p_i (y+z) + p_i^2\, x\, y\, z)$:

$$P[|R \cap S| = t \,/\, |R| = r \text{ et } |S| = s] = \frac{[x^t\, y^r\, z^s]\, f(x,y,z)}{[y^r\, z^s]\, f(1,y,z)}$$

d'où :

$$E[|R \cap S| \,/\, |R| = r \text{ et } |S| = s] = \frac{[y^r\, z^s]\, \frac{\partial f}{\partial x}(1,y,z)}{[y^r\, z^s]\, f(1,y,z)}$$

$$\frac{\partial f}{\partial x}(x,y,z) = f(x,y,z) \times \sum_{i=1}^{d} \frac{p_i^2\, y\, z}{1 + p_i (y+z) + p_i^2\, x\, y\, z}$$

$$\Rightarrow \frac{\partial f}{\partial x}(1,y,z) = f(1,y,z) \cdot \sum_{i=1}^{d} \frac{p_i^2\, y\, z}{(1 + p_i\, y)(1 + p_i\, z)}$$

$$= \sum_{i=1}^{d} [p_i\, y \cdot \prod_{j \neq i} (1 + p_j\, y)] \times [p_i\, z \cdot \prod_{j \neq i} (1 + p_j\, z)]$$

$$= \sum_{i=1}^{d} \left[\prod_{j=1}^{d} (1 + p_j\, y) - \prod_{j \neq i} (1 + p_j\, y) \right] \left[\prod_{j=1}^{d} (1 + p_j z) - \prod_{j \neq i} (1 + p_j z) \right]$$

En posant $\varphi_i(x) = \prod_{j \neq i} (1 + p_j\, x)$

et $\varphi(x) = \prod_{j=1}^{d} (1 + p_j\, x)$

nous avons donc :

$$E[|R \cap S| \,/\, |R| = r \text{ et } |S| = s]$$

$$= \frac{\sum_{i=1}^{d} [x^r]\ (\varphi - \varphi_i) \times [x^s]\ (\varphi - \varphi_i)}{[x^r]\ \varphi \times [x^s]\ \varphi} \quad . \ \blacksquare$$

B - DEPENDANCE FONCTIONNELLE SIMPLE

Hypothèses

Nous supposons ici que R est une relation avec une dépendance fonctionnelle "simple", i.e. qu'il existe une partition des attributs $\{A_1,\ldots,A_k, B_1,\ldots,B_\ell\}$ telle que les seules dépendances fonctionnelles non triviales soient : $B_{i_1} \ldots B_{i_n} \rightarrow A_{j_1} \ldots A_{j_m}$.

Nous nous ramenons alors au cas d'une relation à deux attributs (éventuellement composés) A et B , vérifiant : $B \rightarrow A$, et de domaines respectifs

$$D_A = \prod_{i=1}^{k} D_{A_i} \quad \text{et} \quad D_B = \prod_{j=1}^{\ell} D_{B_j} \ .$$

Notons $d_A = |D_A|$ et $D_B = |d_B|$.

I - LOIS UNIFORMES

La probabilité que $R \cap S$ soit de taille t , sachant que R est de taille r et S de taille s , est :

$$P[|R \cap S| = t \,/\, |R| = r \text{ et } |S| = s] = \frac{\binom{r}{t} \sum_{i=0}^{s-t} \binom{r-t}{i} \binom{d_B - r}{s-t-i} \binom{d_A - r}{d_A}}{\binom{d_B}{s} \; d_A \; t}$$

La taille moyenne de $R \cap S$ est : $\frac{r\,s}{d_A \cdot d_B}$

et la variance :

$$\sigma^2 = \frac{rs}{d_A} \left(1 - \frac{rs}{d_A} + \frac{(r-1)(s-1)}{d_A \; d_B (d_B - 1)}\right) .$$

Démonstration

1) Notons N le nombre de couples (R,S) vérifiant : $|R| = r$ et $|S| = s$, et N_t le nombre de ces couples, qui satisfont de plus à : $|R \cap S| = t$. Alors :

$$P[|R \cap S| = t / |R| = r \text{ et } |S| = s] = \frac{N_t}{N} .$$

. Les valeurs prises par R et S étant indépendantes, N est le produit du nombre de valeurs de R, telles que $R = r$, par le nombre de valeurs de S, telles que $S = s$: $N = \binom{d_B}{r} . d_A{}^r \times \binom{d_B}{s} . d_A{}^s$.

. Pour calculer N_t, nous décomposons la construction d'un couple (R,S) convenant, ainsi :

- choix de $R \cap S$, de taille t : $\binom{d_B}{t} . d_A{}^t$ possibilités ;
- choix de $R \setminus S$, de taille $(r-t)$, et ne reprenant pas les éléments de D_B dans $R \cap S$: $\binom{d_B - t}{r - t} \times d_A{}^{r-t}$ possibilités ;
- choix de $S \setminus R$:
  . nous reprenons j éléments de la colonne B de R $(0 \leqslant j \leqslant r-t)$, avec d'autres éléments de d_A : $\binom{r-t}{j} (d_A - 1)^j$ choix
  . nous complétons S en prenant $(s-t-j)$ éléments de D_B non dans R, pour former les lignes de S : $\binom{d_B - r}{s-t-j} d_A{}^{s-t-j}$ choix.

D'où :

$$N_t = \binom{d_B}{t} d_A^t \times \binom{d_B - t}{r-t} d_A{}^{r-t} \times \sum_{j=0}^{r-t} \binom{r-t}{j} \binom{d_B - r}{s-t-j} (d_A - 1)^j . d_A{}^{s-t-j}$$

$$N_t = d_A{}^{r+s-t} \binom{d_B}{t} \binom{r}{t} \sum_{j=0}^{r-t} \binom{r-t}{j} \binom{d_B - r}{s-t-j} \left(\frac{d_A - 1}{d_A}\right)^j$$

et $P[|R \cap S| = t / |R| = r \text{ et } |S| = s] = \frac{N_t}{N}$

$$= \frac{\binom{r}{t} \sum_{j=0}^{r-t} \binom{r-t}{j} \binom{d_B - r}{s-t-j} \left(\frac{d_A - 1}{d_A}\right)^j}{d_A^t \binom{d_B}{s}} .$$

2) Taille moyenne

$$|R \cap S| = \sum_{\ell \in D_A \times D_B} 1_{(\ell \in R)} \times 1_{(\ell \in S)} \quad .$$

Les variables aléatoires $1_{(\ell \in R)}$ et $1_{(\ell \in S)}$ sont indépendantes. Donc :

$$E(|R \cap S|) = \sum_{\ell \in D_A \times D_B} P(\ell \in R) \times P(\ell \in S) \quad .$$

La loi de probabilité sur $D_A \times D_B$ étant ici uniforme, $P(\ell \in R) = \frac{|R|}{d_A \cdot d_B}$.

Donc : $E(|R \cap S|) = \frac{|R|\ |S|}{d_A\ d_B}$

3) Fonction génératrice

Nous donnons ici la fonction génératrice de $|R \cap S|$, pour $|R| = r$ et $|S| = s$. Cependant, le calcul explicite de la loi de probabilité par cette méthode semble beaucoup plus compliqué que la méthode de dénombrement employée plus haut, bien que l'obtention des moments de la loi soit facile.

Nous représentons les valeurs pouvant être prises par R par :

$$P_R(x_1,\ldots,x_{d_A}, y_1,\ldots,y_{d_B}) = \prod_{i=1}^{d_B} (1 + y_i \sum_{j=1}^{d_A} x_j)$$

$$= \sum_{R_0 \text{ valeur de } R} R_0$$

A S fixé, nous remplaçons, dans l'expression de R , le monôme $x_i y_j$ par x , si $x_i y_j$ n'est pas dans S , et par xy , sinon.

Nous obtenons donc : $x^{|R \cap S|} y^{|R|}$.

En appliquant la transformation ci-dessus à P_R , nous avons :

$$(1 + xy + y(d_A - 1))^s \ . \ (1 + d_A\, y)^{d_B - s}$$

$$= \sum_{r,t} \left[\sum_{\substack{R \text{ tel que } |R| = r \\ \text{et } |R \cap S| = t}} P(R) \right] x^t y^r \quad .$$

Donc : $P[|R \cap S| = t \text{ et } |R| = r / S = S_0] = [x^t y^r]\, \varphi(x,y)$,

en notant $\varphi(x,y) = (1 + d_A\, y)^{d_B - s}\ (1 + xy + y(d_A - 1))^s$.

Nous en tirons :

$$P[|R \cap S| = t / |R| = r \text{ et } S = S_0] = \frac{P[|R \cap S| = t \text{ et } |R| = r / S = S_0]}{P(|R| = r)}$$

et $P[|R \cap S| = t / |R| = r \text{ et } |S| = s]$

$$= \sum_{S_0} P[|R \cap S| = t / |R| = r \text{ et } S = S_0] \ .\ P(S = S_0 / |S| = s)$$

$$= P[|R \cap S| = t / |R| = r \text{ et } S = S_0]$$

$$= \frac{[x^t y^t]\, \varphi(x,y)}{P(|R| = r)}$$

$$= \frac{[x^t y^r] \left\{ \varphi(y,x) \times d_A{}^s . \binom{d_B}{s} \right\}}{(P\,|R| = r).P(|S| = r)}$$

Or $\varphi(x,y).d_A{}^s \binom{d_B}{s} = [z^s] \left\{ (1 + d_A(y+z) + d_A\, y\, z\, (x + d_A - 1))^{d_B} \right\}$.

D'où $P[|R \cap S| = t / |R| = r \text{ et } |S| = s] = \dfrac{[x^t y^r z^s]\ \Phi(x,y,z)}{d_A{}^{r+s} \binom{d_B}{r} \binom{d_B}{s}}$

où $\Phi(x,y,z) = [1 + d_A(y+z) + d_A\, y\, z(x + d_A - 1)]^{d_B}$.

4) Variance

$$\sigma^2 = \frac{[y^r z^s] \frac{\partial^2 \Phi}{\partial x^2}(1,y,z)}{d_A^{r+s} \binom{d_B}{r} \binom{d_B}{s}} + E - E^2$$

$$\frac{\partial^2 \Phi}{\partial x^2}(1,y,z) = d_B(d_B - 1)\,(d_A^2\, y\, z)\quad (1 + d_A\, y)^{d_B-2}\,(1 + d_A\, z)^{d_B-2}$$

$$\Rightarrow \sigma^2 = \frac{rs(r-1)(s-1)}{d_A^2\, d_B(d_B-1)} + \frac{rs}{d_A} - \left(\frac{rs}{d_A}\right)^2$$

$$\sigma^2 = \frac{rs}{d_A}\left[1 - \frac{rs}{d_A} + \frac{(r-1)(s-1)}{d_A\, d_B(d_B-1)}\right] .$$

II - LOIS GENERALES

Notons, pour x_j dans D_A : $P(x_j) = p_j$

et pour y_k dans D_B : $P(y_k) = q_k$.

Loi de $|R \cap S| / |R|$ et $|S|$:

La probabilité que $R \cap S$ soit de taille t, sachant que R est de taille r et S de taille s, est :

$$P[|R \cap S| = t \,/\, |R| = r \text{ et } |S| = s]$$

$$= \frac{[x^r y^t z^s]\, \phi(x,y,z)}{[x^r]\varphi(x).[x^s]\varphi(x)}$$

où $$\phi(x,y,z) = \prod_{k=1}^{d_B} (1 + q_k(x+y) + q_k^2 xz\, [1 + (y-1) \sum_{j=1}^{d_A} p_j^2])$$

et $$\varphi(x) = \prod_{k=1}^{d_B} (1 + q_k\, x) .$$

La taille moyenne de $R \cap S$ est :

$$\frac{\left(\sum_{j=1}^{d_A} p_j^2\right) \sum_{k=1}^{d_B} [x^r]\, (\varphi - \varphi_k)\, [x^s]\, (\varphi - \varphi_k)}{[x^r]\varphi(x).[x^s]\, \varphi(x)}$$

où $$\varphi_k(x) = \prod_{\ell \neq k} (1 + q_\ell\, x) .$$

Démonstration

1) Nous calculerons d'abord, pour une valeur S_0 de S,

$P[|R \cap S| = t \,/\, |R| = r \text{ et } S = S_0]$:

La fonction génératrice des valeurs de R est :

$$F_R(x_1,\ldots,x_{d_A}, y_1,\ldots,y_{d_B}) = \prod_{i=1}^{d_B} \left(\frac{1 + q_i\, y_i \sum_{j=1}^{d_A} p_j\, x_j}{1 + q_i} \right)$$

$$= \sum_R P(R).R \ .$$

Pour obtenir $P[|R \cap S| = t \,/\, |R| = r \text{ et } S = S_0]$, nous remplaçons, dans l'expression de R comme produit de monômes, $x_j\, y_i$ par : xy si $x_j\, y_i$ est aussi dans S, x sinon. Nous obtenons donc, en notant $\varphi_{S_0}(x,y)$ la fonction obtenue par l'application de cette transformation à P_R :

$$\varphi_{S_0}(x,y) = \sum_R P(R).x^{|R|}\, y^{|R \cap S|}$$

$$= \sum_{r,t} \sum_{\substack{R,\ \text{tel que } |R| = r \\ \text{et } |R \cap S| = t}} P(R) \ .\ x^r\, y^t$$

$$= \sum_{r,t} P[|R \cap S| = t \text{ et } |R| = r]\ x^r\, y^t \ .$$

D'autre part, notons $S_0 = \begin{bmatrix} x_{i_1} & y_{j_1} \\ \vdots & \\ x_{i_s} & y_{j_s} \end{bmatrix}$:

$$\varphi_{S_0}(x,y) = \prod_{k \notin \{j_1 \dots j_s\}} \left(\frac{1+q_i\,x}{1+q_i}\right) \cdot \prod_{k=1}^{s} \left(\frac{1+q_{j_k}\,x[1-p_{i_k}+p_{i_k}\,y]}{1+q_{j_k}}\right)$$

$$= \prod_{k=1}^{d_B} \frac{1+q_i\,x}{1+q_i} \cdot \prod_{\ell=1}^{s} \frac{1+q_{j_\ell}(1-p_{i_\ell})\,x + p_{i_\ell}\,q_{j_\ell}\,xy}{1+q_{i_\ell}\,x} \cdot$$

. Par ailleurs, $P[|R \cap S| = t / |R| = r \text{ et } S = S_0] = \dfrac{P(|R \cap S| = t \text{ et } |R| = r / S = S_0)}{P(|R| = r)}$

et $P[|R| = r] = [x^r]\, f_R(1,\dots,1, y,\dots,y)$

$$= [x^r] \prod_{i=1}^{d_B} \frac{1+q_i\,y}{1+q_i}$$

Donc :

$$P[|R \cap S| = t / |R| = r \text{ et } S = S_0] = \frac{[x^r\,y^t]\left\{\prod_{i=1}^{d_B}(1+q_i\,x).\psi_{S_0}(x,y)\right\}}{[x^r]\prod_{i=1}^{d_B}(1+q_i\,x)}$$

où $\psi_{S_0}(x,y) = \prod_{\ell=1}^{s} \dfrac{1+q_{j_\ell}(1-p_{i_\ell})\,x + p_{i_\ell}\,q_{j_\ell}\,xy}{1+q_{i_\ell}\,x}$.

2) $P[|R \cap S| = t / |R| = r \text{ et } |S| = s]$

$$= \sum_{S_0} P(|R \cap S| = t / |R| = r \text{ et } S = S_0).P(S = S_0 / |S| = s)$$

$$. P(S = S_0 / |S| = s) = \frac{\prod_{k=1}^{s} p_{i_k}\,q_{j_k}}{\sum_{\substack{\ell_1 \dots \ell_s \\ m_1 \dots m_s}} p_{\ell_1} \dots p_{\ell_s}\, q_{m_1} \dots q_{m_s}}$$

$$= \frac{\prod_{k=1}^{s} p_{i_k} \cdot q_{j_k}}{[x^s] \prod_{i=1}^{d_B} (1 + q_i\, x)}$$

Donc $P[|R \cap S| = t \,/\, |R| = r \text{ et } |S| = s]$

$$= \frac{\sum_{j_1..j_s \text{ tous distincts}} q_{j_1} \cdots q_{j_s} \sum_{i_1..i_s \text{ non distincts}} p_{i_1} \cdots p_{i_s} [x^r\, y^t] \left\{ \prod_{i=1}^{d_B} (1 + q_i\, x) . \psi_{S_0}(x,y) \right\}}{[x^r] \prod_{i=1}^{d_B} (1 + q_i\, x) \cdot [x^s] \prod_{i=1}^{d_B} (1 + q_i\, x)}$$

Le numérateur ci-dessus vaut :

$$[x^r\, y^t] \left\{ \prod_{i=1}^{d_B} (1+q_i\, x) \sum_{j_1..j_s} q_{j_1} \cdots q_{j_s} \cdot \sum_{i_1 \cdots i_s} \prod_{k=1}^{s} \frac{p_{i_k} (1+q_{j_k} (1-p_{i_k})x + p_{i_k}\, q_{j_k}\, xy}{1 + q_{i_k}\, x} \right.$$

$$= [x^r y^t] \left\{ \prod_{i=1}^{d_B} (1+q_i\, x) \cdot \sum_{j_1 \cdots j_s} \prod_{\ell=1}^{s} \frac{q_{j_\ell}}{1+q_{j_\ell}\, x} \sum_{i_1 \cdots i_s} \prod_{k=1}^{s} p_{i_k} (1 + q_{j_k} (1-p_{i_k})x + p_{i_k}\, q_{j_k}\, x \right.$$

$$\sum_{\substack{i_1 \cdots i_s \\ \text{non distincts}}} \prod_{k=1}^{s} [P_{i_k} (1 + q_{j_k} (1 - p_{i_k})\, x + p_{i_k}\, q_{j_k}\, xy)]$$

$$= \prod_{k=1}^{s} \left[\sum_{j=1}^{d_A} p_j (1 + q_{j_k} (1 - p_j)\, x + p_j\, q_{j_k}\, xy) \right]$$

$$= \prod_{k=1}^{s} \left[\sum_{j=1}^{d_A} \left\{ p_j (1 + q_{j_k}\, x) + p_j{}^2 q_{j_k}\, x(y-1) \right\} \right]$$

$$= \prod_{k=1}^{s} \left[1 + q_{j_k} x + \left(\sum_{j=1}^{d_A} p_j^2 \right) q_{j_k} x(y-1) \right] .$$

Notons $Q = \sum_{j=1}^{d_A} p_j^2$:

$$\sum_{j_1 \ldots j_s} \prod_{k=1}^{s} \frac{q_{j_k}}{1 + q_{j_k} x} \left[1 + q_{j_k} x + Q\, q_{j_k} x(y-1) \right]$$

tous distincts

$$= [z^s] \prod_{k=1}^{d_B} \left[1 + \frac{q_k}{1 + q_k x} \left(1 + q_k x + Q . q_k x(y-1) \right) z \right]$$

$$= \frac{1}{\prod_{k=1}^{d_B} (1 + q_k x)} [z^s] \prod_{k=1}^{d_B} \left[1 + q_k (x+z) + Q\, q_k^2 xz(y-1) \right] .$$

D'où :

$$P[|R \cap S| = t \,/\, |R| = r \text{ et } |S| = s]$$

$$= \frac{[x^r y^t z^s] \prod_{k=1}^{d_B} \left(1 + q_k (x+z) + Q\, q_k^2 xz(y-1) \right)}{[x^r] \left\{ \prod_{k=1}^{d_B} (1 + q_k x) \right\} . [x^s] \left\{ \prod_{k=1}^{d_B} (1 + q_k x) \right\}} .$$

3) Taille moyenne

Notons $\phi(x,y,z)$ le produit

$$\prod_{k=1}^{d_B} (1 + q_k(x+z) + q_k^2 xz\,[1 + (y-1)Q])$$

$$E(\,|R \cap S| \,/\, |R| = r \text{ et } |S| = s) = \frac{\sum_{t \geqslant 0} t.[x^r y^t z^s]\,\phi(x,y,z)}{[x^r] \prod_i (1+p_i x).[x^s] \prod_i (1+p_i x)}$$

$$= \frac{[x^r z^s]\,\frac{\partial \phi}{\partial y}(x,1,z)}{[x^r] \prod_i (1+p_i x)\;[x^s] \prod_i (1+p_i x)}$$

$$\frac{\partial \phi}{\partial y}(x,y,z) = \phi(x,y,z).\sum_{k=1}^{d_B} \frac{q_k^2 xz\,Q}{1 + q_k(x+z) + q_k^2 xz(1 + (y-1)Q)}$$

Donc :

$$\frac{\partial \phi}{\partial y}(x,1,z) = \phi(x,1,z).\sum_{k=1}^{d_B} \frac{q_k^2 xz\,Q}{(1+q_k x)\;(1+q_k z)}$$

$$= Q.\sum_{k=1}^{d_B} q_k^2 xz \prod_{\ell \neq k} (1+q_\ell x)\;(1+q_\ell z)$$

et $$[x^r z^s]\,\frac{\partial \phi}{\partial y}(x,1,z) = Q.\sum_{k=1}^{d_B} [x^r](\varphi - \varphi_k).[x^s](\varphi - \varphi_k)$$

où $\varphi(x) = \prod_{\ell=1}^{d_B} (1 + q_\ell x)$ et $\varphi_k(x) = \prod_{\ell \neq k} (1 + q_\ell x)$

Nous en tirons :

$$E[|R \cap S| \,/\, |R| = r \text{ et } |S| = s] = \frac{Q \cdot \sum_{k=1}^{d_B} [x^r]\,(\varphi - \varphi_k) \cdot [x^s]\,(\varphi - \varphi_k)}{[x^r]\,\varphi \qquad [x^s]\,\varphi} \quad .$$

CHAPITRE IV

PROJECTION

Nous étudions dans ce chapitre la taille de la projection d'une relation de taille donnée, dans le cas d'un schéma libre, et dans le cas d'un schéma à clé "simple". Nous nous plaçons successivement dans l'hypothèse où les lois sur les attributs sont uniformes, et dans le cas général. Nous obtenons la loi de la taille de la projection, connaissant la taille de la relation initiale, et en déduisons la taille moyenne et la variance.

A - RELATION SOUS DEPENDANCE FONCTIONNELLE

Soit $R(A_1,\ldots,A_n)$ un schéma de relation, sous dépendance fonctionnelle. Nous notons $\Pi_{A_{i_1} \ldots A_{i_k}}(R)$ la projection de R sur les colonnes $A_{i_1} \ldots A_{i_k}$ $(1 \leqslant k \leqslant n)$

Pour simplifier les notations, nous nous ramenons au cas d'une relation à deux attributs A et B, en posant :

$$A = A_{i_1} \ldots A_{i_k} \qquad \text{et} \qquad D_A = D_{A_{i_1}} x \ldots x D_{A_{i_k}}$$

$$B = A_{j_1} \ldots A_{j_{n-k}} \qquad \text{et} \qquad D_B = D_{B_{j_1}} x \ldots x D_{B_{j_{n-k}}}$$

où $\{ j_1 \ldots j_{n-k} \} \cup \{ i_1, - i_k \} = \{ 1, -, n \}$

Dans le cas où la relation initiale avait plus de deux attributs, les attributs A et B ne sont pas simples, mais ils sont indépendants.

Nous avons donc à chercher la loi de la taille de $\Pi_A(R)$, R étant une relation à deux attributs indépendants A et B.

1) LOIS UNIFORMES

Nous supposons ici que les lois de probabilités sur les domaines D_A et D_B sont uniformes, i.e. :

$$\forall\ a \in D_A,\quad P(a) = \frac{1}{d_A} \qquad \text{où} \qquad d_A = |D_A|$$

$$\forall\ b \in D_B,\quad P(b) = \frac{1}{d_B} \qquad \text{où} \qquad d_B = |D_B|$$

Loi de $|\pi_A(R)|$, sachant $|R|$:

$$P[|\pi_A(R)| = t \ / \ |R| = r] = \frac{\binom{d_A}{t}}{\binom{d_A d_B}{r}} \cdot \sum_{j=0}^{t} (-1)^{t+j} \binom{t}{j} \binom{j d_B}{r}$$

démonstration :

Nous donnerons deux démonstrations de ce résultat, l'une étant combinatoire et l'autre utilisant les fonctions génératrices.

1) les lois sur D_A et D_B étant uniformes, la loi de probabilité sur l'ensemble des valeurs de R, de taille donnée, l'est aussi

$$P\ [R \text{ prend la valeur } R_o \ / \ |R| = r] = \frac{1}{\binom{d_A d_B}{r}}$$

En notant $N_{t,r}$ le nombre de valeurs de R qui satisfont à : $|\pi_A(R)| = t$ et $|R| = r$, la probabilité $P[\pi_A(R)| = t/\ |R| = r]$ est donc égale à : $N_{t,r} \times P\ [R \text{ prend la valeur } Ro/\ |R| = r]$

i.e. : $$P[|\pi_A(R)| = t \ / \ |R| = r] = \frac{N_{t,r}}{\binom{d_A d_B}{r}}$$

Calcul de $N_{t,r}$:

Pour obtenir une valeur de R, telle que $|R| = r$ et $|\Pi_A(R)| = t$, nous pouvons choisir d'abord les t éléments de $\Pi_A(R)$, puis compléter cette projection en une valeur acceptable de R. Pour cela, nous reprenons n_i fois chaque élément de $\Pi_A(R)$ $(1 \leqslant i \leqslant t)$, et $n_1 + .. + n_t = r$, pour en faire n_i éléments distincts (a, b_j) de R. Il y a donc $\prod_{i=1}^{t} \binom{d_B}{n_i}$ façons de procéder, pour $(n_1, \ldots, n_t)$ fixé. Finalement:

$$N_{t,r} = \binom{d_A}{t} \sum_{\substack{n_1 \ldots n_t \leqslant 1 \\ n_1 + .. n_t = r}} \prod_{i=1}^{t} \binom{d_B}{n_i} .$$

Le nombre de termes de la somme ci-dessus, étant le nombre de partitions de r en t sommants, est équivalent, pour r et t assez grands, à : $\frac{r^{t-1}}{t!(t-1)!}$. Il est donc utile de trouver une expression plus simple de $N_{t,r}$:

$$\sum_{\substack{n_1 \ldots n_t \geqslant 1 \\ n_1 + \ldots + n_t = r}} \prod_{t=1}^{t} \binom{d_B}{n_i} = \sum_{j=1}^{t} (-1)^{t+j} \binom{t}{j} \binom{Jd_B}{r}$$

(pour la démonstration de cette formule, voir l'appendice I)

Nous avons donc :

$$P[\, |\Pi_A(R)| = t / |R| = r] = \frac{\binom{d_A}{t}}{\binom{d_A d_B}{r}} \times \sum_{j=1}^{t} (-1)^{t+j} \binom{t}{j} \binom{Jd_B}{r}$$

2) FONCTIONS GENERATRICES :

Soit P le polynôme formel $\prod_{\ell \in D_A \times D_B} (1+x_\ell)$:

$$P = \sum_{J \geqslant 0} \sum_{\ell_1 \ldots \ell_j} x_{\ell_1} \ldots x_{\ell_j}$$

A chaque valeur $(\ell_1 \ldots \ell_j)$ de R, nous associons le monôme $x_{\ell_1} \ldots x_{\ell_j}$

P représente donc l'ensemble des valeurs, de taille quelconque, pouvant être prises par R. De plus, en remplaçant chaque x_ℓ par x, nous obtenons comme coefficient de x^j, le nombre de valeurs de R de taille j :

$$P(x) = (1+x)^{d_A d_B} = \sum_{j=0}^{d_A d_B} \binom{d_A d_B}{j} x^J.$$

Pour obtenir le nombre $N_{t,r}$ de valeurs de R, de taille r et dont la projection sur A est de taille t, nous remplaçons d'abord chaque x_ℓ par $x_{\Pi(\ell)}$. Ensuite, dans les monômes $x_{\Pi(\ell 1)}^{i_1} \ldots x_{\Pi(\ell j)}^{i_j}$, nous remplaçons chaque $x_{\Pi(\ell)}^{i}$ par : $\alpha . x^i$. Soit $Q(\alpha, x)$ le polynôme ainsi obtenu :

$$Q(\alpha,x) = \sum_{\substack{t \geqslant 0 \\ r \geqslant 0}} N_{t,r} \, \alpha^t . x^r,$$

$N_{t,r}$ est le coefficient de $\alpha^t x^r$ dans $Q(\alpha,x)$.

EXPRESSION DE $Q(\alpha,x)$:

Nous avons d'abord remplacé $P = \prod_{\ell \in D_A \times D_B} (1+x_\ell)$

par: $\prod_{a \in D_A} (1+x_a)^{d_B}$.

Ensuite, la substitution de $\alpha.x^i$ à $x^i_{\Pi(\ell)}$ peut se faire sur chaque

polynôme $(1+x_a)^{d_B}$: $(1+x_a)^{d_B} = 1 + [(1+x_a)^{d_B} - 1]$

le polynôme, à valuation nulle, $(1+x_a)^{d_B}-1$ est donc transformé en : $\alpha[(1+x)^{d_B}-1.]$.

Finalement :

$$Q(\alpha,x) = \left(1+\alpha[(1+x)^{d_B}-1]\right)^{d_A}$$

$$= \sum_{t=0}^{d_A} \binom{d_A}{t} . \alpha^t . [(1+x)^{d_B}-1]^t$$

$$[(1+x)^{d_B}-1]^t = \sum_{j=0}^{t} \binom{t}{j} (-1)^{t-j} (1+x)^{jd_B}$$

$$= \sum_{\substack{0 \leqslant j \leqslant t \\ 0 \leqslant r \leqslant jd_B}} . \binom{t}{j}(-1)^{t-j} \binom{jd_B}{r} x^r$$

Donc :

$$Q(\alpha,x) = \sum_{\substack{0 \leqslant j \leqslant t \leqslant d_A \\ 0 \leqslant r \leqslant j\, d_B}} (-1)^{t-j} \binom{d_A}{t} \binom{t}{j} \binom{jd_B}{r} \alpha^t x^r$$

$$= \sum_{\substack{0 \leq t \leq d_A \\ 0 \leq r \leq d_A d_B}} \binom{d_A}{t} \left[\sum_{j=0}^{t} (-1)^{t-j} \binom{t}{j} \binom{jd_B}{r}\right] \alpha^t x^r$$

Nous en tirons :

$$N_{t,r} = \binom{d_A}{t} \sum_{j=0}^{t} (-1)^{t-j} \binom{t}{j} \binom{jd_B}{r}$$

et :

$$P\left[|\pi_A(R)| = t / |R| = r\right] = \frac{\binom{d_A}{t}}{\binom{d_A d_B}{r}} \sum_{j=0}^{t} (-1)^{t-j} \binom{t}{j} \binom{jd_B}{r}$$

Taille moyenne et variance de $|\pi_A(R)|$:

La taille moyenne de $|\pi_A(R)|$, sachant que $|R| = r$, est :

$$E\left[|\pi_A(R)| / |R| = r\right] = d_A \left[1 - \frac{\binom{d_A d_B - d_B}{r}}{\binom{d_A d_B}{r}}\right]$$

$$\simeq r\left[1 - \frac{r-1}{d_A}\right]$$

La variance est :

$$\sigma^2\left[|\Pi_A(R) / |R| = r\right] = d_A^2 \left[\frac{\binom{d_A d_B - 2d_B}{r}}{\binom{d_A d_B}{r}} - \left(\frac{\binom{d_A d_B - d_B}{r}}{\binom{d_A d_B}{r}}\right)^2\right]$$

$$+ d_A \left[\frac{\binom{d_A d_B - d_B}{r}}{\binom{d_A d_B}{r}} - \frac{\binom{d_A d_B - 2d_B}{r}}{\binom{d_A d_B}{r}}\right]$$

$$\simeq \frac{r(r-1)}{2d_A}$$

Démonstration :

a/- Taille moyenne

$$E\left[|\pi_A(R)| / |R| = r\right] = \sum_{t=0}^{r} t.P\left[|\pi_A(R)| = t / |R| = r\right]$$

$$= \frac{1}{\binom{d_A d_B}{r}} \sum_{t=0}^{r} t.\, N_{t,r}$$

$$\text{Or : } Q(\alpha,x) = \sum_{r,t} N_{t,r}\, \alpha^t x^r$$

$$\text{donc : } \frac{\partial Q}{\partial \alpha}(\alpha,x) = \sum_r \left[\sum_t t.N_{t,r}\, \alpha^{t-1}\right] x^r$$

$$\text{et } \frac{\partial Q}{\partial \alpha}(1,x) = \sum_r \left[\sum_t t N_{t,r}\right] x^r \ .$$

$$\text{Donc : } E[|\Pi_A(R)| \ / \ |R| = r] = \frac{1}{\binom{d_A\ d_B}{r}} \times \text{coefficient de } x^r \text{ dans } \frac{\partial Q}{\partial \alpha}(1,x)$$

$$\frac{\partial Q}{\partial \alpha}(\alpha,x) = \frac{\partial}{\partial \alpha}\left[(1+\alpha[(1+x)^{d_B} - 1])^{d_A}\right]$$

$$= d_A.(1+\alpha[(1+x)^{d_B} - 1])^{d_A - 1}\ [(1+x)^{d_B} - 1]$$

$$\frac{\partial Q}{\partial \alpha}(1,x) = d_A.(1+x)^{d_A d_B - d_B}\ [(1+x)^{d_B} - 1]$$

$$= d_A[(1+x)^{d_A d_B} - (1+x)^{d_A d_B - d_B}] \ .$$

Nous en tirons :

$$E[|\Pi_A(R)|/|R|=r] = \frac{d_A}{\binom{d_A\ d_B}{r}} \times \left[\binom{d_A\ d_B}{r} - \binom{d_A d_B - d_B}{r}\right]$$

$$= d_A\left[1 - \frac{\binom{d_A d_B - d_B}{r}}{\binom{d_A\ d_B}{r}}\right]$$

La formule approchée s'obtient en remarquant que, pour $d \geqslant r$: $\binom{d}{r} \simeq \frac{d^r}{r!}$. En reportant dans l'expression de la taille moyenne, nous obtenons :

$$E[|\Pi_A(R)| \;/\; |R| = r] \simeq d_A\left[1 - \left(\frac{d_A d_B - d_B}{d_A d_B}\right)^r\right]$$

$$\simeq d_A\left[1 - \left(1 - \frac{1}{d_A}\right)^r\right]$$

$$\simeq d_A\left[\frac{r}{d_A} - \frac{r(r-1)}{2\,d_A}\right]$$

$$\simeq r\left[1 - \frac{r-1}{2\,d_A}\right] \quad . \quad \blacksquare$$

b/ Variance

$$\sigma^2[|\Pi_A(R)| \;/\; |R| = r] = E[|\Pi_A(R)|^2 \;/\; |R| = r] - (E[|\Pi_A(R)| \;/\; |R| = r])^2$$

$$. \; E[|\Pi_A(R)| \;/\; |R| = r] = d_A\left[1 - \frac{\binom{d_A d_B - d_B}{r}}{\binom{d_A\; d_B}{r}}\right]$$

$$. \; E[|\Pi_A(R)|^2 \;/\; |R| = r] = \sum_{t=0}^{r} t^2 \, P[|\Pi_A(R)| = t / |R| = r]$$

$$= \frac{1}{\binom{d_A\; d_B}{r}} \sum_{t=0}^{r} t^2 \, \tilde{N}_{t,r}$$

$$= \frac{1}{\binom{d_A\; d_B}{r}} \left[\sum_{t=0}^{r} t(t-1)\, N_{t,r} + \sum_{t=0}^{r} t\, N_{t,r}\right]$$

$$= E[|\Pi_A(R)| / |R| = r] + \frac{1}{\binom{d_A\ d_B}{r}} \sum_{t=0}^{r} t(t-1)\ N_{t,r}$$

$$\frac{\partial^2 Q}{\partial \alpha^2}(\alpha,x) = \sum_r \left[\sum_t t(t-1)\ N_{t,r}\ \alpha^{t-2}\right] x^r$$

Donc : $\sum_t t(t-1)\ N_{t,r}$ est le coefficient de x^r dans $\frac{\partial^2 Q}{\partial \alpha^2}(1,x)$.

$$\cdot\ \frac{\partial^2 Q}{\partial \alpha^2}(\alpha,x) = d_A(d_A-1)(1+\alpha[(1+x)^{d_B}-1])^{d_A-2}[(1+x)^{d_B}-1]^2$$

$$\frac{\partial^2 Q}{\partial \alpha^2}(1,x) = d_A(d_A-1)(1+x)^{d_A d_B-2d_B}[(1+x)^{d_B}-1]^2$$

$$= d_A(d_A-1)[(1+x)^{d_A d_B} - 2(1+x)^{d_A d_B - d_B} + (1+x)^{d_A d_B - 2d_B}]$$

D'où :

$$\sum_t t(t-1)\ N_{t,r} = d_A(d_A-1)\left[\binom{d_A\ d_B}{r} - 2\binom{d_A d_B - d_B}{r} + \binom{d_A d_B - 2d_B}{r}\right].$$

Finalement :

$$\sigma^2[|\Pi_A(R)| / |R| = r] =$$

$$d_A\left[1 - \frac{\binom{d_A d_B - d_B}{r}}{\binom{d_A\ d_B}{r}}\right] + d_A(d_A-1)\left[\frac{\binom{d_A d_B - 2d_B}{r}}{\binom{d_A\ d_B}{r}} - 2\frac{\binom{d_A d_B - d_B}{r}}{\binom{d_A\ d_B}{r}} + 1\right]$$

$$- d_A^2\left[1 - 2\frac{\binom{d_A d_B - d_B}{r}}{\binom{d_A\ d_B}{r}} + \frac{\binom{d_A d_B - d_B}{r}^2}{\binom{d_A\ d_B}{r}}\right]$$

$$= d_A^2\left[\frac{\binom{d_A d_B - 2d_B}{r}}{\binom{d_A\ d_B}{r}} - \frac{\binom{d_A d_B - d_B}{r}^2}{\binom{d_A\ d_B}{r}^2}\right] + d_A\left[\frac{\binom{d_A d_B - d_B}{r}}{\binom{d_A\ d_B}{r}} - \frac{\binom{d_A d_B - 2d_B}{r}}{\binom{d_A\ d_B}{r}}\right]$$

Valeur approchée

$$\frac{\binom{d_A d_B - d_B}{r}}{\binom{d_A\ d_B}{r}} = \left(1 - \frac{1}{d_A}\right)^r \cdot \frac{\left(1 - \frac{1}{d_A d_B - d_B}\right)\cdots\left(1 - \frac{r-1}{d_A d_B - d_B}\right)}{\left(1 - \frac{1}{d_A d_B}\right)\cdots\left(1 - \frac{r-1}{d_A\ d_B}\right)}$$

$$\simeq \left(1 - \frac{1}{d_A}\right)^r \left(1 - \frac{r(r-1)}{1(d_A d_B - d_B)}\right)\left(1 + \frac{r(r-1)}{2d_A\ d_B}\right)$$

$$\simeq \left(1 - \frac{1}{d_A}\right)^r \left(1 - \frac{r(r-1)}{2d_A\ d_B} - \frac{r(r-1)}{2d_A^2 d_B}\right)\left(1 + \frac{r(r-1)}{2d_A\ d_B}\right)$$

$$\simeq \left(1 - \frac{r}{d_A} + \frac{r(r-1)}{2d_A^2} - \frac{r(r-1)(r-2)}{6d_A^3}\right)\left(1 - \frac{r(r-1)}{2d_A^2 d_B}\right)$$

$$\simeq 1 - \frac{r}{d_A} + \frac{r(r-1)}{2\ d_A^2} - \frac{r(r-1)(r-2)}{6\ d_A^3} - \frac{r(r-1)}{2d_A^2 d_B}$$

En élevant au carré, nous obtenons :

$$\frac{\binom{d_A d_B - d_B}{r}^2}{\binom{d_A\ d_B}{r}^2} = 1 - \frac{2r}{d_A} + \frac{r(2r-1)}{d_A^2} - \frac{r(r-1)}{d_A^2 d_B} - \frac{r(r-1)(4r-2)}{3\ d_A^3}\ .$$

Un calcul analogue au précédent nous donne :

$$\frac{\binom{d_A d_B - 2d_B}{r}}{\binom{d_A\ d_B}{r}} = 1 - \frac{2r}{d_A} + \frac{2r(r-1)}{d_A^2} - \frac{4r(r-1)(r-2)}{3\,d_A^3} - \frac{r(r-1)}{d_A^2 d_B} \quad .$$

Si nous reportons les valeurs approchées dans l'expression de la variance, nous obtenons :

$$\sigma^2[|\Pi_A(R)| \ / \ |R| = r] \simeq d_A^2\left[-\frac{r}{d_A^2} + \frac{2r(r-1)}{d_A^3}\right]$$

$$+ d_A\left[\frac{r}{d_A} - \frac{3\,r(r-1)}{2\,d_A^2}\right]$$

$$\simeq \frac{r(r-1)}{2\,d_A} \quad . \quad \blacksquare$$

2) LOIS GENERALES

Loi de $[|\Pi_A(R)| \ / \ |R|$

$$P[|\Pi_A(R)| = t \ / \ |R| = r] =$$

$$\frac{[\alpha^t x^r] \prod_{i=1}^{d_A} \{1 + \alpha[\prod_{j=1}^{d_B}(1 + p_i q_j x) - 1]\}}{[x^r] \prod_{i=1}^{d_A} \prod_{j=1}^{d_B}(1 + p_i q_j x)}$$

$$= \frac{\sum_{k=0}^{t} (-1)^{t-k} \binom{d_A - k}{t - k} a_{k,r}}{[x^r] \prod_{i,j} (1 + p_i q_j x)}$$

où $a_{k,r}$ est le coefficient de x^r dans :

$$\sum_{i_1 \ldots i_k} \prod_{u=1}^{k} \prod_{j=1}^{d_B} (1 + p_{i_u} q_j x)$$

Démonstration

A chaque valeur $R_0 = (\ell_1 \ldots \ell_r)$ pouvant être prise par R, nous associons le monôme $p(\ell_1)\ldots p(\ell_r).\ell_1 \ldots \ell_r$. L'ensemble des valeurs possibles pour R est alors représenté par le polynôme :

$$P(\ell_1, \ldots, \ell_{d_A \times d_B}) = \prod_{\ell \in D_A \times D_B} (1 + p(\ell).\ell)$$

$$= \sum_r \sum_{i_1 \ldots i_r} P(\ell_{i_1}) \ldots p(\ell_{i_r}).\ell_{i_1} \ldots \ell_{i_r} \quad .$$

Pour obtenir le polynôme générateur des valeurs de $R_1 : Q(x_{\ell_1}, \ldots)$, nous divisons $P(x_{\ell_1}, \ldots)$ par $P(1, \ldots, 1)$:

$$Q(x_{\ell_1}, \ldots) = \prod_{\ell \in D_A \times D_B} \left[\frac{1 + p(\ell).x_\ell}{1 + p(\ell)} \right]$$

$$= \sum_{\substack{R_0 \text{ valeur de } R \\ (R_0 = x_{\ell_{i_1}} \ldots x_{\ell_{i_{|R_0|}}})}} P(R_0).R_0 \quad .$$

Si maintenant nous divisons une ligne ℓ en deux parties x_i, y_j : $x_i \in D_A$, $y_j \in D_B$:

$$p(\ell) = p(x_i).p(y_j) = p_i.q_j \ .$$

Donc nous obtenons pour polynôme générateur de R :

$$Q(x_1,\ldots x_{d_A} \ , \ y_1,\ldots y_{d_B}) = \prod_{i=1}^{d_A} \prod_{j=1}^{d_B} \left(\frac{1 + p_i \ q_j \ x_i \ y_j}{1 + p_i \ q_j} \right)$$

$$= \sum_{R_0} P(R_0) \ . \ R_0 \ .$$

Pour obtenir les valeurs R_0 de R , telles que : $|R_0| = r$ et $|\Pi_A(R_0)| = t$:

. Nous remplaçons chaque y_j par 1 , et regroupons les x_i égaux.

. puis chaque $x_i^{n_i}$ est remplacé par $\alpha.x^{n_i}$.

Notons f_{α,x_i} cette transformation.

Nous obtenons ainsi :

$$f_{\alpha,x_1} \circ \ldots \circ f_{\alpha,x_{d_A}} \ [Q(x_1,\ldots,x_{d_A} \ , \ 1,\ldots,1)]$$

$$= \sum_{\substack{0 \leqslant r \leqslant d_A d_B \\ 0 \leqslant t \leqslant r}} \ \sum_{\substack{R_0 \text{ valeur de R,} \\ \text{telle que : } |R_0| = r \\ |\Pi_A(R_0)| = t}} P(R_0) \ . \ \alpha^t \ x^r$$

$$= \sum_{r,t} P[|\Pi_A(R)| = t \ \text{ et } \ |R| = r] \ \alpha^t \ x^r \ .$$

. $P[|\Pi_A(R)| = t \ / \ |R| = r] =$

$$P[|\Pi_A(R)| = t \text{ et } |R| = r] \ / \ P[|R| = r] \ .$$

. Par ailleurs : $P[|R| = r]$ est le coefficient de x^r dans

$$Q(x,\ldots,x\ ,\ 1,\ldots,1) = \prod_{i,j} \left(\frac{1 + p_i\ q_j\ x}{1 + p_i\ q_j}\right) \ .$$

Donc : $P[|\Pi_A(R)| = t \ / \ |R| = r]$

$$= \frac{[\alpha^t\ x^r]\ f_{\alpha,x_1} \circ \ldots \circ f_{\alpha,x_{d_A}} \left[\prod_{i,j} \left(\frac{1 + p_i\ q_j\ x_i}{1 + p_i\ q_j}\right)\right]}{[x^r] \prod_{i,j} \left(\frac{1 + p_i\ q_j\ x}{1 + p_i\ q_j}\right)} \ .$$

D'où, en simplifiant par $\prod_{i,j} (1 + p_i\ q_j)$: $P[|\Pi_A(R)| = t \ / \ |R| = r]$

$$= \frac{[\alpha^t x^r]\ f_{\alpha,x_1} \circ \ldots \circ f_{\alpha,x_{d_A}} [\prod_{i,j} (1 + p_i\ q_j\ x_i)]}{[x^r] \prod_{i,j} (1 + p_i\ q_j\ x)} \ .$$

Or : $f_{\alpha,x_1} \circ \ldots \circ f_{\alpha,x_{d_A}} [\prod_{i,j} (1 + p_i\ q_j\ x_i)]$

$$= \prod_{i=1}^{d_A} \{f_{\alpha,x_i} [\prod_{j=1}^{d_B} (1 + p_i\ q_j\ x_i)]\} \ .$$

Comme $f_{\alpha,x_i} [P(x_i)] = P(0) + \alpha[P(x) - P(0)]$, nous obtenons :

$$f_{\alpha,x_1} \circ \ldots \circ f_{\alpha,x_{d_A}} [\prod_{i,j} (1 + p_i\ q_j\ x_i)]$$

$$= \prod_{i=1}^{d_A} \{1 + \alpha\ [\prod_{j} (1 + p_i\ q_j\ x) - 1]\} \ ,$$

D'où la première expression de la probabilité.

Pour obtenir la seconde, nous remarquons que :

$$f_{\alpha,x_1} \circ \ldots \circ f_{\alpha,x_{d_A}}(P) = \sum_{k=0}^{d_A} P_k(x).\alpha^k.(1-\alpha)^{d_A-k} \quad ,$$

où $P_k(x)$ est le polynôme obtenu en remplaçant, dans P, k des variables $x_1 \ldots x_{d_A}$ par x, et les autres par 0. (cf. Appendice II).

En notant :

$$\sum_{i_1 \ldots i_k} \prod_{u=1}^{k} \prod_{j=1}^{d_B} (1 + p_{i_u} q_j x) = \sum_{r=0}^{kd_B} a_{k,r} x^r \quad ,$$

nous obtenons :

$$f_{\alpha,x_1} \circ \ldots \circ f_{\alpha,x_{d_A}} \left[\prod_{i,j} (1 + p_i q_j x_i) \right]$$

$$= \sum_{k,r} a_{k,r} x^r . \alpha^k (1-\alpha)^{d_A-k}$$

$$= \sum_{r,t} \left[\sum_{k} a_{k,r} \binom{d_A - k}{t-k} (-1)^{t-k} \right] \alpha^t x^r$$

D'où le deuxième résultat. ∎

Taille moyenne

$$E[|\Pi_A(R)| \;/\; |R| = r] = d_A - \frac{[x^r] \sum_{i=1}^{d_A} \prod_{k \neq i} \prod_{j=1}^{d_B} (1 + p_k q_j x)}{[x^r] \prod_{i,j} (1 + p_i q_j x)} .$$

Démonstration

$$P[|\Pi_A(R)| = t \;/\; |R| = r] = \frac{\varphi(t)}{P[|R| = r]}$$

Donc :

$$E[|\Pi_A(R)| \;/\; |R| = r] = \frac{1}{P[|R| = r]} \sum_{t=1}^{r} t.\,\varphi(t)$$

$$= \frac{E(\varphi)}{P[|R| = r]}$$

$\varphi(t) = [\alpha^t\, x^r]\, f(\alpha,x)$.

Donc : $E(\varphi) = [x^r] \frac{\partial f}{\partial \alpha}(1,x)$.

(Démonstration identique à celle faite dans le cas de lois uniformes).

$$f(\alpha,x) = \prod_{i=1}^{d_A} \{1 + \alpha[\prod_{j=1}^{d_B} (1 + p_i\, q_j\, x) - 1]\} .$$

Donc :

$$\frac{\partial f}{\partial \alpha}(\alpha,x) = f(\alpha,x) \times \sum_{i=1}^{d_A} \frac{\prod_{j=1}^{d_B} (1 + p_i\, q_j\, x) - 1}{1 + \alpha[\prod_{j=1}^{d_B} (1 + p_i\, q_j\, x) - 1]}$$

$$\frac{\partial f}{\partial \alpha}(1,x) = f(1,x) \cdot \sum_{i=1}^{d_A} \frac{\prod_{j=1}^{d_B} (1 + p_i\, q_j\, x) - 1}{\prod_{j=1}^{d_B} (1 + p_i\, q_j\, x)}$$

$$= \sum_{i=1}^{d_A} \left[\prod_{j=1}^{d_B} (1 + p_i\, q_j\, x) - 1 \right] \times \prod_{k \neq i} \prod_{j=1}^{d_B} (1 + p_k\, q_j\, x)$$

$$\frac{\partial f}{\partial \alpha}(1,x) = d_A\, f(1,x) - \sum_{i=1}^{d_A} \prod_{k \neq i} \prod_{j=1}^{d_B} (1 + p_k\, q_j\, x) \; .$$

Remarquons que : $P[|R| = r] = [x^r]\, f(1,x)$.

Donc :

$$E[|\Pi_A(R)| \;/\; |R| = r]$$

$$= \frac{[x^r]\left[d_A . f(1,x) - \sum_{i=1}^{d_A} \prod_{k \neq i} \prod_{j=1}^{d_B} (1 + p_k\, q_j\, x)\right]}{[x^r]\, f(1,x)}$$

$$= d_A - \frac{[x^r] \sum_{i=1}^{d_A} \prod_{k \neq i} \prod_{j=1}^{d_B} (1 + p_k\, q_j\, x)}{[x^r] \prod_{i,j} (1 + p_i\, q_j\, x)} \; . \quad \blacksquare$$

B - RELATION AVEC UNE DÉPENDANCE FONCTIONNELLE "SIMPLE"

Nous appelons "relation avec une dépendance fonctionnelle simple" une relation telle qu'il existe une partition des attributs en deux ensembles $\{A_1,\ldots,A_k\}$ et $\{B_1,\ldots,B_j\}$, et que les seules dépendances fonctionnelles non triviales soient :

$$B_{i_1}\ldots B_{i_n} \rightarrow A_{j_1}\ldots A_{j_n}$$

Nous pouvons donc nous ramener au cas d'une relation R à deux attributs A et B vérifiant : $B \rightarrow A$. Nous avons alors :

$$|\Pi_B(R)| = |R| \text{. (évident)}$$

$$1 \leqslant |\Pi_A(R)| \leqslant |R|$$

I - LOIS UNIFORMES

Nous supposons ici que les lois de probabilité sur les domaines D_A et D_B sont uniformes. Alors : la loi de probabilité sur l'ensemble des valeurs de R , de taille donnée, l'est aussi :

$$P[R \text{ prend la valeur } R_0 \;/\; |R| = r] = \frac{1}{\binom{d_B}{r} d_A{}^r} \quad .$$

LOI de $|\Pi_A(R)| \;/\; |R| = r$:

$$P[|\Pi_A(R)| = t \;/\; |R| = r] = \frac{\binom{d_A}{t}}{d_A{}^r} \cdot \sum_{j=0}^{t} (-1)^{t+j} \binom{t}{j} j^r$$

Démonstration

Une démonstration de ce résultat fait appel au dénombrement et est d'esprit proche de la première démonstration du cas sans dépendance fonctionnelle. Elle est exposée en détail dans [14].

Nous donnerons ici une démonstration faisant appel aux fonctions génératrices.

Soit P le polynôme formel :

$$(x_1 + \ldots + x_{d_A})^r = \sum_{t_1 \ldots t_r} x_{t_1} \ldots x_{t_r} ,$$

dans cette dernière écriture les x_t ne sont pas nécessairement distincts, mais sont ordonnés : chaque monôme $x_{t_1} \ldots x_{t_r}$ correspond à une colonne $(t_1 \ldots t_r)$ de l'attribut A, donc à $\binom{d_B}{r}$ valeurs de R.

Pour obtenir le nombre N_t de valeurs de A, donnant une projection sur A de taille t, nous regroupons d'abord tous les x_n égaux dans chaque monôme $x_{i_1} \ldots x_{i_r}$. Nous obtenons donc des monômes : $x_{i_1}^{n_1} \ldots x_{i_t}^{n_t}$ $(1 \leqslant n_1, \ldots, n_t$ et $n_1 + \ldots + n_t = r)$.

Puis nous remplaçons successivement chaque x_i^n par : $\alpha . x^n$. En notant f_{α, x_i} cette transformation, nous obtenons :

$$f_{\alpha, x_1} \circ \ldots \circ f_{\alpha, x_{d_A}} (P) = Q(\alpha, x) :$$

$$Q(\alpha, x) = \sum_{t=1}^{r} \frac{N_t}{\binom{d_B}{r}} \alpha^t x^r .$$

Par ailleurs :

$$f_{\alpha,x_1} \circ \ldots \circ f_{\alpha,x_{d_A}} (P) = \sum_{j=0}^{d_A} \binom{d_A}{j} P(\underbrace{x,\ldots,x}_{j \text{ fois}} ; 0,\ldots,0)\alpha^j (1-\alpha)^{d_A-j}$$

(cf. Appendice II).

Donc :

$$Q(\alpha,x) = \sum_{j=0}^{d_A} \binom{d_A}{j} (jx)^r \alpha^j (1-\alpha)^{d_A-j}$$

$$= \sum_{j=0}^{d_A} \binom{d_A}{j} j^r \alpha^j \sum_{u=0}^{d_A-j} \binom{d_A-j}{u} (-1)^u \alpha^u \cdot x^r$$

$$= \sum_{t=0}^{d_A} \binom{d_A}{t} \left[\sum_{j=0}^{t} (-1)^{t-j} \binom{t}{j} j^r\right] \alpha^t x^r .$$

D'où : $N_t = \binom{d_A}{t} \binom{d_B}{r} \sum_{j=0}^{t} (-1)^{t-j} \binom{t}{j} j^r$

$P[|\Pi_A(R)| = t \ / \ |R| = r] = N_t \times P[R \text{ prend valeur } R_0 \ / \ |R| = r]$

$$= N_t \times \frac{1}{\binom{d_B}{r} d_A^r}$$

$$= \frac{\binom{d_A}{t}}{d_A^r} \cdot \sum_{j=0}^{t} (-1)^{t-j} \binom{t}{j} j^r . \quad \blacksquare$$

Taille moyenne et variance de $|\Pi_A(R)|$:

$$E[|\Pi_A(R)| \ / \ |R| = r] = d_A\left[1 - \left(1 - \frac{1}{d_A}\right)^r\right]$$

$$\simeq r\left[1 - \frac{r-1}{2\,d_A}\right]$$

$$\sigma^2[|\Pi_A(R)| \ / \ |R| = r] = d_A(d_A - 1)\left(1 - \frac{2}{d_A}\right)^r$$

$$+ d_A\left(1 - \frac{1}{d_A}\right)^r - d_A^2\left(1 - \frac{1}{d_A}\right)^{2r}$$

$$\simeq \frac{r(r-1)}{2\,d_A} \quad .$$

Démonstration

a) Taille moyenne

$$Q(\alpha,1) = \sum_t d_A{}^r . P[|\Pi_A(R)| = t \ / \ |R| = r] \ . \alpha^t$$

Donc :

$$\sum_t P[|\Pi_A(R)| = t \ / \ |R| = r] \ . \ \alpha^t = d_A{}^{-r} \ . \ Q(\alpha,1) \ .$$

Soit φ la fonction génératrice de $|\Pi_A(R)|$:

$$\varphi(\alpha) = \sum_t P[|\Pi_A(R)| = t \ / \ |R| = r] \ . \ \alpha^t$$

$$= \frac{1}{d_A{}^r} \ Q(\alpha,1) \ .$$

Donc : $E[|\Pi_A(R)| \;/\; |R| = r] = \varphi'(1) = \frac{1}{d_A{}^r} \frac{\partial}{\partial \alpha} Q(1,1)$

$$. \; Q(\alpha,1) = \sum_{j=0}^{d_A} \binom{d_A}{j} j^r \; . \; \alpha^j (1-\alpha)^{d_A - j}$$

$$. \; \frac{\partial Q}{\partial \alpha}(1,1) = d_A{}^{r+1} - d_A(d_A - 1)^r$$

$$= d_A{}^{r+1} \left[1 - \left(\frac{d_A - 1}{d_A}\right)^r\right] \quad .$$

Donc :

$$E[|\Pi_A(R)| \;/\; |R| = r] = d_A \left[1 - \left(\frac{d_A - 1}{d_A}\right)^r\right] \quad .$$

b) Variance

$$\sigma^2 = \varphi''(1) + \varphi'(1) - \varphi'^2(1)$$

$$\varphi''(1) = d_A(d_A - 1) \left[\left(1 - \frac{2}{d_A}\right)^r - 2\left(1 - \frac{1}{d_A}\right)^r + 1\right]$$

Donc :

$$\sigma^2[|\Pi_A(R)| \;/\; |R| = r] = d_A(d_A - 1) \left(1 - \frac{2}{d_A}\right)^r + d_A \left(1 - \frac{1}{d_A}\right)^r - d_A^2 \left(1 - \frac{1}{d_A}\right)^{2r} .$$

II - LOIS GENERALES

Les lois de probabilité sur D_A et D_B sont ici quelconques. Nous montrerons d'abord que la loi de probabilité sur D_B n'influe pas sur la taille de $\Pi_A(R)$. Puis nous calculerons la loi de $|\Pi_A(R)|$.

Non-influence de la loi de D_B :

$$P[|\Pi_A(R)| = t \ / \ |R| = r \text{ et } B \text{ prend une valeur } B_0]$$

$$= P[|\Pi_A(R)| = t \ / \ |R| = r]$$

En effet : pour une ligne donnée, la valeur prise par A ne dépend de la valeur prise par B que dans la mesure où b est déjà dans R , un couple (a,b) ne pouvant être répété dans la relation. L'attribut B étant ici clé de R , les valeurs prises par A à chaque ligne, sont donc indépendantes des valeurs prises par B .

Nous supposons donc la loi sur D_A quelconque, et :

$$P(x_i) = p_i \qquad (1 \leqslant i \leqslant d_A).$$

Loi de la taille de $\Pi_A(R)$:

$$P[|\Pi_A(R)| = t \ / \ |R| = r]$$

$$= \sum_{j=1}^{t} (-1)^{t-j} \binom{d_A - j}{t - j} \sum_{1 \leqslant i_1 < \ldots < i_j \leqslant d_A} (p_{i_1} + \ldots + p_{i_j})^r \quad .$$

Démonstration

A chaque valeur de l'attribut $A : x_{i_1} \ldots x_{i_r}$, nous associons le polynôme $p_{i_1} \ldots p_{i_r} \cdot x_{i_1} \ldots x_{i_r}$. L'ensemble des colonnes A possibles, et de leur probabilité, est donc décrit par le polynôme formel :

$$P(x_1, \ldots, x_{d_A}) = \sum_{i_1 \ldots i_r} P_{i_1} \ldots p_{i_r} \, x_{i_1} \ldots x_{i_r}$$

$$= (p_1 x_1 + \ldots + p_{d_A} x_{d_A})^r \quad .$$

La fonction génératrice de $|\Pi_A(R)|$ s'obtient, comme dans le cas de lois uniformes, par :

$$\varphi(\alpha) = \sum_{t \geqslant 0} P[|\Pi_A(R)| = t \ / \ |R| = r] \times \alpha^t$$

$$= f_{\alpha, x_1} \circ \ldots \circ f_{\alpha, x_{d_A}} (P)(\alpha, 1) \quad .$$

Or : $f_{\alpha, x_1} \circ \ldots \circ f_{\alpha, x_{d_A}} (P) = \sum_{j=0}^{d_A} P_j(x) . \alpha^j (1 - \alpha)^{d_A - j}$

où $P_j(x) = \sum_{\sigma \in S_{d_A}} P[y_{\sigma(1)}, \ldots, y_{\sigma(d_A)}]$

$(y_1 = \ldots = y_j = x \ ; \ y_{j+1} = \ldots = y_{d_A} = 0)$

[S_{d_A} est ici le groupe des permutations à d_A éléments].

Donc :

$$\varphi(\alpha) = \sum_{j=0}^{d_A} P_j(1).\alpha^j.(1-\alpha)^{d_A-j}$$

$$= \sum_t \left[\sum_j (-1)^{t-j} P_j(1) \binom{d_A - j}{t-j} \right] \alpha^t \ .$$

Il suffit alors d'écrire explicitement $P_j(1)$ pour obtenir le résultat . ■

Taille moyenne et variance de $|\Pi_A(R)|$

$$E[|\Pi_A(R)| \ / \ |R| = r] = d_A - \sum_{j=1}^{d_A} (1-p_j)^r$$

$$\sigma^2[|\Pi_A(R)| \ / \ |R| = r] = \sum_{j=1}^{d_A} (1-p_j)^r - \left[\sum_{j=1}^{d_A} (1-p_j)^r \right]^2$$

$$+ 2 \sum_{1 \leqslant i < j \leqslant d_A} (1 - p_i - p_j)^r \ .$$

Démonstration

$$E[|\Pi_A(R)| \ / \ |R| = r] = \varphi'(1)$$

$$\sigma^2[|\Pi_A(R)| \ / \ |R| = r] = \varphi''(1) + \varphi'(1) - \varphi'^2(1) \ .$$

Il suffit donc de différencier $\varphi(\alpha) = \sum_j P_j(1) \ \alpha^j (1-\alpha)^{d_A-j}$ pour obtenir le résultat. ■

CHAPITRE V

EQUIJOINTURE

Nous étudions ici l'équijoin de deux relations R et S sur un attribut commun. Nous supposons que R a deux attributs A et B, S deux attributs A et C, de domaines respectifs D_A, D_B et D_C, et que les hypothèses suivantes sont respectées :

a) Il n'y a pas de dépendance fonctionnelle dans R et dans S.

B) R et S prennent des valeurs indépendantes. Donc : $P((R,S)) = P(R) \times P(S)$.

Nous cherchons la loi de taille de $R \bowtie S$, en supposant connues les tailles de R et S, ou seulement la taille d'une relation. Nous donnons la fonction génératrice de $|R \bowtie S|$, ainsi que la taille moyenne et la variance, dans le cas de lois uniformes puis dans le cas général.

A - LOIS UNIFORMES

1) Hypothèses

* Nous supposons que les lois sur D_A , D_B et D_C sont uniformes, i.e. : $\forall\, a \in D_A$, $P(a) = \frac{1}{d_A}$ etc.

  D'où, pour tout triplet (a, b, c) : $P(abc) = \frac{1}{d_A\, d_B\, d_C}$.

* Pour une taille donnée, les valeurs d'une relation R , ou S , ou $R \times S$, sont donc équiprobables.

  Par exemple :

  Soit R_0 une valeur de R , de taille r .

$$P[R = R_0 \;/\; |R| = r] = \frac{1}{\binom{d_A\, d_B}{r}}$$

* Nous supposerons de plus que, quelle que soit leur taille, toutes les valeurs pouvant être prises par R (resp. S) sont équiprobables.

  En termes de fonctions génératrices, ceci revient à associer à R le polynôme formel

$$P_R = \prod_{\substack{a \in D_A \\ b \in D_B}} (1 + x_a\, y_b)$$

$$= \sum_{\substack{a_1 \ldots a_k \in D_A \\ b_1 \ldots b_k \in D_A}} (x_{a_1}\, y_{b_1}) \times \ldots \times (x_{a_k}\, y_{b_k}) \; .$$

Remarque

Il n'y a pas bijection entre les monômes de P_R et les valeurs pouvant être prises par R . Par exemple, les valeurs $\begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix}$ et $\begin{pmatrix} a_1 & b_2 \\ a_2 & b_1 \end{pmatrix}$ sont associées au même monôme $(x_{a_1} x_{a_2} y_{b_1} y_{b_2})$.

2) Fonction génératrice de $|R \bowtie S|$, connaissant R(A)

Notons $R(A) = x_1^{j_1} \ldots x_{d_A}^{j_{d_A}}$, où $j_i (1 \leqslant i \leqslant T)$ est le nombre de fois où x_i est présent dans R(A) : $j_i \geqslant 0$, $j_1 + \ldots + j_{d_A} = |R|$.

Alors :

La fonction génératrice de $|R \bowtie S|$, connaissant R(A), est :

$$\varphi(x) = 2^{-d_A d_C} \prod_{t=1}^{d_A} (1 + x^{j_t})^{d_C}$$

En particulier :

* $P[|R \bowtie S| = r \ / \ R(A)]$

$$= 2^{-d_A d_C} \sum_{\substack{i_1 \ldots i_{d_A} \geqslant 0 \\ i_1 j_1 + \ldots + i_{d_A} j_{d_A} = r}} \binom{d_C}{i_1} \ldots \binom{d_C}{i_{d_A}}$$

* $E[|R \bowtie S| \ / \ R(A)] = \frac{1}{2} d_C . |R|$.

* $\sigma^2[|R \bowtie S| \ / \ R(A)] = \frac{1}{4} d_C \sum_{t=1}^{d_A} j_t^2$

Démonstration

a) Fixer $R(A)$ revient à fixer $(j_1,\ldots,j_{d_A})$.

Pour les valeurs de R correspondant à $R(A)$ fixé, la taille du join avec S sera la même : cette taille ne dépend en effet que des colonnes $R(A)$ et $S(A)$; si nous associons à $S(A)$ le T-uplet $(i_1,\ldots,i_{d_A})$: $|R \bowtie S| = \sum_{t=1}^{d_A} i_t \cdot j_t$.

Considérons donc :

$$P_S(x^{j_1},\ldots,x^{j_{d_A}},1,\ldots,1) = \prod_{t=1}^{d_A} (1+x^{j_t})^{d_C} = \phi(x)$$

ϕ s'obtient en remplaçant dans P_S tous les z_C par 1 , et les x_t par x^{j_t} .

Le coefficient de x^r dans ϕ est alors le nombre de valeurs de S , dont le join avec une valeur R_0 de R , ayant une colonne A fixée ci-dessus, est de taille r .

Pour avoir la fonction génératrice de la taille de $R_0 \bowtie S$, il suffit de diviser par $\phi(1)$: nombre total de valeurs de S possibles.

D'où : $\varphi(x) = \frac{\phi(x)}{\phi(1)}$

$$= 2^{-d_A d_C} \prod_{t=1}^{d_A} (1+x^{j_t})^{d_C} \quad .$$

b) En développant $\varphi(x)$, nous obtenons les probabilités $P[|R \bowtie S| = r \ / \ R(A)]$, comme coefficient de x^r :

$$\varphi(x) = 2^{-d_A d_C} P_S(x^{j_1}, \ldots, x^{j_{d_A}}, 1, \ldots, 1)$$

$$P_S(x_1, \ldots, x_{d_A}, 1, \ldots, 1) = \sum_{i_1 \ldots i_{d_A} \geqslant 0} N(i_1 \ldots i_{d_A}) x_1^{i_1} \ldots x_{d_A}^{i_{d_A}}$$

où $N(i_1 \ldots i_{d_A})$ est le nombre de tableaux S , de colonne A comprenant i_k fois a_k $(1 \leqslant k \leqslant d_A , i_1 \ldots i_{d_A} \geqslant 0)$

$$N(i_1 \ldots i_{d_A}) = \binom{d_C}{i_1} \times \ldots \times \binom{d_C}{i_{d_A}} .$$

Donc :

$$P_S(x^{j_1}, \ldots, x^{j_{d_A}}, 1, \ldots, 1) = \sum_{i_1 \ldots i_{d_A} \geqslant 0} \binom{d_C}{i_1} \ldots \binom{d_C}{i_{d_A}} x^{i_1 j_1 + \ldots + i_{d_A} j_{d_A}}$$

et

$$\varphi(x) = \sum_{i_1 \ldots i_{d_A} \geqslant 0} 2^{-d_A d_C} \binom{d_C}{i_1} \ldots \binom{d_C}{i_{d_A}} x^{i_1 j_1 + \ldots + i_{d_A} j_{d_A}}$$

$$= \sum_{r=0}^{d_A} \left[\sum_{\substack{i_1 \ldots i_{d_A} \geqslant 0 \\ i_1 j_1 + \ldots + i_{d_A} j_{d_A} = r}} 2^{-d_A d_C} \binom{d_C}{i_1} \ldots \binom{d_C}{i_{d_A}} \right] x^r .$$

D'où :

$$P[|R \bowtie S| = r \ / \ R(A)] = \sum_{\substack{i_1 \ldots i_{d_A} \geqslant 0 \\ i_1 j_1 + \ldots + i_{d_A} j_{d_A} = r}} 2^{-d_A d_C} \binom{d_C}{i_1} \ldots \binom{d_C}{i_{d_A}} .$$

c) La taille moyenne de $R \bowtie S$ est donnée par : $\varphi'(1)$, et la variance par : $\varphi''(1) + \varphi'(1) - \varphi'^2(1)$.

$$\varphi(x) = 2^{-d_A d_C} \prod_{t=1}^{d_A} (1 + x^{j_t})^{d_C}$$

$$\Rightarrow \frac{\varphi'(x)}{\varphi(x)} = d_C \sum_{t=1}^{d_A} \frac{\frac{d}{dx}(1 + x^{j_t})}{1 + x^{j_t}} = d_C \sum_{t=1}^{d_A} \frac{j_t \, x^{j_t - 1}}{1 + x^{j_t}} \quad .$$

D'où :

$$\varphi'(1) = \varphi(1) \; d_C \sum_{t=1}^{d_A} \frac{j_t}{2} \quad .$$

Comme $\sum_t j_t = |R|$, et $\varphi(1) = 1$:

la taille moyenne de $R \bowtie S$ est donc : $\frac{1}{2} \; d_C \; |R|$.

Pour obtenir la variance, nous calculons $\varphi''(1)$:

$$\varphi'(x) = \varphi(x) \left[d_C \sum_t \frac{j_t \, x^{j_t - 1}}{1 + x^{j_t}} \right]$$

$$\Rightarrow \varphi''(x) = \varphi'(x) \left[d_C \sum_t \frac{j_t \, x^{j_t - 1}}{1 + x^{j_t}} \right]$$

$$+ \varphi(x) \; d_C \sum_t \left(\frac{j_t (j_t - 1) x^{j_t - 2}}{1 + x^{j_t}} - \frac{j_t^2 \, x^{2j_t - 2}}{(1 + x^{j_t})^2} \right)$$

$$\varphi''(x) = \varphi(x)\left[\left(d_C \sum_t \frac{j_t\, x^{j_t - 1}}{1 + x^{j_t}}\right)^2 + d_C \sum_t \frac{j_t(j_t - 1)x^{j_t - 2}}{1 + x^{j_t}} - d_C \sum_t \frac{j_t^2\, x^{2J_t - 2}}{(1 + x^{j_t})^2}\right]$$

$$\varphi''(1) = [\varphi'(1)]^2 + \frac{d_C}{2} \sum_t j_t(j_t - 1) - \frac{d_C}{4} \sum_t (j_t)^2$$

$$\varphi''(1) = [\varphi'(1)]^2 + \frac{d_C}{4} \sum_t (j_t)^2 - \varphi'(1) \quad .$$

D'où :

$$\sigma^2 = \frac{d_C}{4} \sum_t (j_t)^2 \quad . \quad \blacksquare$$

3) Fonction génératrice de $|R \bowtie S|$, connaissant $|R| = r$

La fonction génératrice de $|R \bowtie S|$, connaissant $|R| = r$, est :

$$\psi(x) = 2^{-d_A d_C} \sum_{\substack{j_1 \ldots j_{d_A} \geqslant 0 \\ j_1 + \ldots + j_{d_A} = r}} P[R(A) = (x_1^{j_1} \ldots) / |R| = r] \times \prod_{t=1}^{d_A} (1 + x^{j_t})^{d_C}$$

$$= \sum_{t \geqslant 0} \left[\sum_{\substack{i_1 \ldots \geqslant 0 \\ j_1 \ldots \geqslant 0 \\ i_1 j_1 + \ldots + i_{d_A} j_{d_A} = t \\ j_1 + \ldots + j_{d_A} = r}} \frac{2^{-d_A d_C}}{\binom{d_A\, d_B}{r}} \binom{d_B}{j_1} \ldots \binom{d_B}{j_{d_A}} \binom{d_C}{i_1} \ldots \binom{d_C}{i_{d_A}} \right] x^t \quad .$$

Démonstration

$$P[|R \bowtie S| = t \,/\, |R| = r] = \sum_{\substack{j_1 \ldots \geqslant 0 \\ j_1 + \ldots = r}} P[|R \bowtie S| = t \,/\, R(A) = (x_1{}^{j_1} \ldots)]$$

$$\times \; P[R(A) = (x_1{}^{j_1} \ldots) / |R| = r] \; .$$

La fonction génératrice de $|R \bowtie S|$, sachant $|R|$, est donc :

$$\psi(x) = \sum_{t \geqslant 0} P[|R \bowtie S| = t \,/\, |R| = r] . x^t$$

$$= \sum_{t \geqslant 0} x^t \sum_{\substack{j_1 \ldots \geqslant 0 \\ j_1 + \ldots = r}} P[|R \bowtie S| = t \,/\, R(A) = (x_1{}^{j_1} \ldots)]$$

$$\times \; P[R(A) = (x_1{}^{j_1} \ldots) \,/\, |R| = r]$$

$$= \sum_{\substack{j_1 \ldots \geqslant 0 \\ j_1 + \ldots = r}} P[R(A) = (x_1{}^{j_1} \ldots) \,/\, |R| = r] \sum_{t \geqslant 0} x^t \; P[|R \bowtie S| = t \,/\, R(A) =$$

Nous avons montré plus haut que la fonction génératrice de $|R \bowtie S| \,/\, R(A) = (x_1{}^{j_1} \ldots)$ est :

$$\sum_{r \geqslant 0} P[|R \bowtie S| = r \,/\, R(A) = (x_1{}^{j_1} \ldots)] = 2^{-d_A d_C} \prod_{t=1}^{d_A} (1 + x^{j_t})^{d_C} \; .$$

Donc :

$$\psi(x) = \sum_{\substack{j_1 \cdots \; j_{d_A} \geqslant 0 \\ j_1 + \ldots + j_{d_A} = |R|}} P[R(A) = (x_1{}^{j_1} \ldots) \,/\, |R|] . 2^{-d_A d_C} \prod_{t=1}^{d_A} (1 + x^{j_t})^{d_C} \; .$$

Nous avons donc à calculer $P[R(A) = (x_1^{j_1}\ldots) \;/\; |R|]$. Comme nous avons supposé l'équiprobabilité des valeurs de R , c'est :

$$\frac{N_{j_1 \ldots j_{d_A}}}{N_{|R|}} \quad , \text{ où}$$

. $N_{|R|}$ est le nombre de valeurs de R , de taille $|R|$ fixée

. $N_{j_1 \ldots j_{d_A}}$ est le nombre de telles valeurs, telles que :

$$R(A) = (x_1^{j_1} \ldots x_{d_A}^{j_{d_A}})$$

. $N_{j_1 \ldots j_{d_A}} = \binom{d_B}{j_1} \ldots \binom{d_B}{j_{d_A}}$

. $N_{|R|} = \binom{d_A\ d_B}{|R|}$

Donc :

$$P[R(A) = (x_1^{j_1}\ldots) \;/\; |R|] = \frac{\prod_{t=1}^{d_A} \binom{d_B}{j_t}}{\binom{d_A\ d_B}{|R|}} \quad .$$

D'où :

$$\psi(x) = \sum_{\substack{j_1 \ldots \geqslant 0 \\ j_1 + \ldots = |R|}} \frac{2^{-d_A d_C} \prod_{t=1}^{d_A} \binom{d_B}{j_t}}{\binom{d_A\ d_B}{|R|}} \prod_{t=1}^{d_A} (1 + x^{j_t})^{d_C} \quad .$$

En développant $\prod_{t=1}^{d_A} (1 + x^{j_t})^{d_C}$, nous obtenons :

$$\psi(x) = \sum_{t \geqslant 0} \Big[\sum_{\substack{i_1 \ldots \geqslant 0 \\ j_1 \ldots \geqslant 0 \\ J_1 + \ldots = r \\ i_1 j_1 + \ldots = t}} \frac{2^{-d_A d_C}}{\binom{d_A\ d_B}{r}} \prod_{t=1}^{d_A} \binom{d_B}{j_t} \binom{d_C}{i_t} \Big] x^t$$

4) Probabilités, taille moyenne et variance

$$P[|R \bowtie S| = t \ / \ |R| = r]$$

$$= \sum_{\substack{i_1 \ldots \geqslant 0 \\ j_1 \ldots \geqslant 0 \\ J_1 + \ldots = r \\ i_1 j_1 + \ldots = t}} \frac{2^{-d_A d_C}}{\binom{d_A\ d_B}{r}} \prod_{t=1}^{d_A} \binom{d_B}{j_t} \binom{d_C}{i_t}$$

$$E(|R \bowtie S| \ / \ |R| = r) = \frac{1}{2} r\, d_C$$

$$\sigma^2(|R \bowtie S| \ / \ |R| = r) = \frac{1}{4} \frac{d_C}{(d_A d_B - 1)} .r[(r-1)(d_B - 1) + d_A d_B - 1]$$

$$\simeq \frac{1}{4} r\, d_C \left(1 + \frac{r-1}{d_A}\right)$$

Démonstration

a) Calcul de la taille moyenne

$$E(|R \bowtie S| \ / \ |R| = r) = \sum_{\substack{j_1 \ldots \geqslant 0 \\ j_1 + \ldots = r}} \sum_{t \geqslant 0} P[|R \bowtie S| = t \ / \ R(A) = (x^{j_1} \ldots)] x^t$$

$$\times\ P[R(A) = (x^{j_1} \ldots) / |R| = r]$$

$$= \sum_{\substack{j_1 \ldots \geq 0 \\ j_1 + \ldots = r}} P(R(A) = (x^{j_1} \ldots)/|R| = r) \times E(|R \bowtie S|/R(A) = (x^{j_1} \ldots)$$

$$= \frac{1}{2} r\, d_C \sum_{\substack{j_1 \ldots \geq 0 \\ j_1 + \ldots = r}} P(R(A) = (x^{j_1} \ldots)/|R| = r)$$

$$= \frac{1}{2} r\, d_C \quad .$$

b) Calcul de la variance

$$\sigma^2[|R \bowtie S|/|R| = r] = \psi''(1) + \psi'(1) - \psi'^2(1) \ .$$

Notons $\varphi_{(j_1 \ldots j_{d_A})}(x) = 2^{-d_A d_C} \prod_{t=1}^{d_A} (1 + x^{j_t})^{d_C}$

$$\psi(x) = \sum_{\substack{j_1 \geq 0 \\ j_1 + \ldots = r}} P[R(A) = (x_1^{j_1} \ldots)/|R| = r].\varphi_{(j_1 \ldots)}(x)$$

$$\Rightarrow \psi''(x) = \sum_{(j_1 \ldots)} P[R(A) = (x_1^{j_1} \ldots)/|R| = r].\varphi''_{(j_1 \ldots)}(x)$$

$$\Rightarrow \psi''(1) = \sum_{(j_1 \ldots)} P[R(A) = (x^{j_1} \ldots)/|R| = r].\varphi''_{(j_1 \ldots)}(1) \quad .$$

Or (cf. plus haut) :

$$\varphi''_{(j_1 \ldots)}(1) = \frac{1}{4} d_C \sum_t (j_t)^2 + \frac{1}{4} d_C^2 \ r^2 - \frac{1}{2} d_C \ r$$

$$\Rightarrow \psi''(1) = \frac{1}{4} d_C \sum_{(j_1 \ldots)} \sum_{t=1}^{d_A} j_t^2 \; P[R(A) = (x_1^{j_1} \ldots)/|R| = r]$$

$$+ \left(\frac{1}{4} d_C^2 r^2 - \frac{1}{2} d_C \, r \right) \sum_{(j_1 \ldots)} P[R(A) = (x_1^{j_1} \ldots)/|R| = r]$$

$$\psi''(1) = \frac{1}{4} d_C \sum_{(j_1 \ldots)} \sum_{t=1}^{d_A} j_t^2 \; P[R(A) = (x_1^{j_1} \ldots)/|R| = r]$$

$$+ \frac{1}{4} d_C^2 r^2 - \frac{1}{2} d_C \, r \quad .$$

D'où :

$$\sigma^2[|R \bowtie S|/|R| = r] = \frac{1}{4} d_C \sum_{\substack{(j_1 \geqslant 0) \\ j_1 + \ldots = r}} \sum_{t} j_t^2 P[R(A) = (x_1^{j_1} \ldots)/|R| = r] \; .$$

Nous avons vu*

$$P[R(A) = (x_1^{j_1} \ldots) / |R| = r] = \frac{1}{\binom{d_A \; d_B}{r}} \prod_{t=1}^{d_A} \binom{d_B}{j_t}$$

Donc :

$$\sigma^2[|R \bowtie S|/|R| = r] = \frac{d_C}{4\binom{d_A \; d_B}{r}} \sum_{\substack{j_1 \ldots \; 0 \\ j_1 + \ldots = r}} \sum_{s=1}^{d_A} j_s^2 \prod_{t=1}^{d_A} \binom{d_B}{j_t} \; .$$

Pour calculer $\sum_{\substack{j_1,\ldots\geq 0 \\ j_1+\ldots=r}} \left(\sum_{s=1}^{d_A} j_s^2\right)\left(\prod_{t=1}^{d_A}\binom{d_B}{j_t}\right)$,

nous utilisons $P_R(x_1,\ldots,x_{d_A},1,\ldots,1) = \prod_{t=1}^{d_A}(1+x_t)^{d_B}$, noté par abréviation dans la suite $P_R(x_1,\ldots,x_{d_A})$.

$$P_R(x_1,\ldots,x_{d_A}) = \sum_{r\geq 0} \sum_{\substack{j_1\ldots\geq 0 \\ j_1+\ldots=r}} \binom{d_B}{j_1}\ldots\binom{d_B}{j_{d_A}} x_1^{j_1}\ldots x_{d_A}^{j_{d_A}}$$

$$\frac{\partial P_R}{\partial x_s}(x_1,\ldots,x_{d_A}) = \sum_{r\geq 0} \sum_{\substack{j_1\ldots\geq 0 \\ j_1+\ldots=r}} j_s \binom{d_B}{j_1}\ldots\binom{d_B}{j_{d_A}} x_1^{j_1}\ldots x_s^{j_s-1}\ldots$$

$$\Rightarrow \sum_{s=1}^{d_A}\frac{\partial P_R}{\partial x_s}(x,\ldots,x) = \sum_{r\geq 0}\left[\sum_{\substack{j_1\ldots\geq 0 \\ j_1+\ldots=r}} \sum_{s=1}^{d_A} j_s \binom{d_B}{j_1}\ldots\binom{d_B}{j_{d_A}}\right]x^{r-1}$$

$$= \sum_{r\geq 0} r\binom{d_A\, d_B}{r}.x^{r-1}$$

$$\frac{\partial^2 P_R}{\partial x_s^2}(x_1,\ldots,x_{d_A}) = \sum_{r\geq 0} \sum_{\substack{j_1\ldots j_{d_A}\geq 0 \\ j_1+\ldots+j_{d_A}=r}} j_s(j_s-1)\prod_{t=1}^{d_A}\binom{d_B}{j_t} x_1^{j_1}..x_s^{j_s-2}\ldots$$

$$\sum_{s=1}^{d_A}\frac{\partial^2 P_R}{\partial x_s^2}(x_1,\ldots,x_{d_A})$$

$$= \sum_{r\geq 0} \sum_{\substack{j_1\ldots\geq 0 \\ j_1+\ldots=r}} \sum_{s=1}^{d_A} j_s^2 \prod_{t=1}^{d_A}\binom{d_B}{j_t} x_1^{j_1}\ldots x_s^{j_s-2}\ldots x_{d_A}^{j_{d_A}}$$

$$- \sum_{r\geq 0} \sum_{\substack{j_1\ldots\geq 0 \\ j_1+\ldots=r}} \sum_{s=1}^{d_A} j_s \prod_{t=1}^{d_A}\binom{d_B}{j_t} x_1^{j_1}\ldots x_s^{j_s-2}\ldots x_{d_A}^{j_{d_A}}$$

D'où :

$$(1)\quad \left\{ \begin{array}{l} \displaystyle\sum_{r \geqslant 0} \Bigg[\sum_{\substack{j_1 \ldots \geqslant 0 \\ j_1 + \ldots = r}} \sum_{s=1}^{d_A} j_s^2 \prod_{t=1}^{d_A} \binom{d_B}{j_t} \Bigg] x^{r-2} \\ \qquad = \displaystyle\sum_{s=1}^{d_A} \frac{\partial^2 P_R}{\partial x_s^2}(x, \ldots, x) + \sum_{r \geqslant 0} r \binom{d_A\ d_B}{r} x^{r-2} \end{array} \right.$$

$$\frac{\partial^2 P_R}{\partial x_s^2}(x_1, \ldots, x_{d_A}) = \frac{d_B(d_B-1)}{(1+x_s)^2} P_R(x_1, \ldots, x_{d_A}) \ .$$

Donc :

$$\sum_{s=1}^{d_A} \frac{\partial^2 P_R}{\partial x_s^2}(x, \ldots, x) = d_B(d_B-1) \times d_A (1+x)^{d_A d_B - 2}$$

$$= d_A d_B (d_B - 1) \sum_{r=0}^{d_A d_B - 2} \binom{d_A\ d_B - 2}{r} x^r$$

En identifiant les coefficients de x^r à droite et à gauche de l'équation (1), nous obtenons :

$$\sum_{\substack{j_1 \ldots \geqslant 0 \\ j_1 + \ldots = r}} \sum_{s=1}^{d_A} j_s^2 \prod_{t=1}^{d_A} \binom{d_B}{j_t}$$

$$= d_A\ d_B (d_B - 1) \binom{d_A\ d_B - 2}{r-2} + r \binom{d_A\ d_B}{r}$$

$$= r \binom{d_A\ d_B}{r} \left[1 + (r-1) \frac{d_B - 1}{d_A d_B - 1} \right]$$

D'où

$$\sigma^2[|R\bowtie S|/|R| = r] = \frac{1}{4} d_C\, r\left[1 + (r-1)\frac{(d_B - 1)}{d_A d_B - 1}\right]$$

$$\simeq \frac{1}{4} r\, d_C\left(1 + \frac{(r-1)}{d_A}\right) . \qquad \blacksquare$$

5) Loi de $|R \bowtie S| / R = R_0$ et $|S| = s$

La probabilité que $R \bowtie S$ soit de taille t, sachant que R prend la valeur R_0, et que S est de taille s, est :

$$P[|R\bowtie S| = t / R = R_0 \text{ et } |S| = s]$$

$$= \frac{1}{\binom{d_A d_C}{s}} [x^t y^s] \prod_{i=1}^{d_A} (1 + x^{j_i} y)^{d_C}$$

où j_t est le nombre de fois où x_t apparaît dans R_0.

La taille moyenne de $R\bowtie S$ est alors :

$$E[|R\bowtie S| / R = R_0 \text{ et } |S| = s] = \frac{|R_0| s}{d_A}$$

et la variance :

$$\sigma^2 = \frac{|R_0|^2 s(s-1) d_C}{d_A(d_A d_C - 1)} - \frac{|R_0|^2 s^2}{d_A} + \sum_{i=1}^{d_A} j_i^2 \cdot \frac{s(d_A d_C - s)}{d_A(d_A d_C - 1)}$$

Si r et s sont petits par rapport à d_A et d_C :

$$\sigma^2 \simeq \frac{\left(\sum_i j_i^2\right) s}{d_A} \left(1 - \frac{r^2 s}{d_A \cdot \sum_i j_i^2}\right)$$

Démonstration

a) Soit $R_0 = x_1^{j_1} \ldots x_{d_A}^{j_{d_A}}$

$$P_S(x^{j_1}, \ldots, x^{j_{d_A}}, y, \ldots, y) = \sum_S x^{|R \times S|} . y^{|S|}$$

$$= \sum_{s,t} N_{t,s}\, x^t\, y^s$$

où $N_{t,s}$ est le nombre de valeurs de S, pour lesquelles $|R \bowtie S| = t$ et $|S| = s$

Nous en tirons :

$$P[|R \bowtie S| = t / R = R_0 \text{ et } |S| = s]$$

$$= P[|R \bowtie S| = t \text{ et } |S| = s / R = R_0] \times P[|S| = s]^{-1}$$

$$= \frac{N_{t,s}}{2^{d_A d_C}} \times \left[\frac{\binom{d_A d_C}{s}}{2^{d_A d_C}}\right]^{-1} = \frac{N_{t,s}}{\binom{d_A d_C}{s}}$$

$$= \frac{[x^t y^s] P_S(x^{j_1}, \ldots, x^{j_{d_A}}, y, \ldots, y)}{\binom{d_A d_C}{s}}$$

$$= \frac{1}{\binom{d_A d_C}{s}} [x^t y^s] \prod_{i=1}^{d_A} (1 + x^{j_i} y)$$

Nous en déduisons une écriture explicite de cette probabilité :

$$P[|R \bowtie S| = t \,/\, R = R_0 \text{ et } |S| = s]$$

$$= \frac{1}{\binom{d_A\, d_C}{s}} \times \sum_{\substack{\ell_1 \ldots \ell_{d_A} \geqslant 0 \\ \ell_1 + \ldots + \ell_{d_A} = s \\ \ell_1 j_1 + \ldots + \ell_{d_A} j_{d_A} = t}} \binom{d_C}{\ell_1} \ldots \binom{d_C}{\ell_{d_A}} .$$

b) Taille moyenne de $R \bowtie S$

Remarquons que :

$$\sum_{t \geqslant 0} t.[x^t y^s]\, f(x,y) = [y^s]\, \frac{\partial f}{\partial x}(1,y)$$

donc en notant $\prod_{i=1}^{d_A} (1 + x^{j_i} y)^{d_C} = f(x,y)$:

$$E[|R \bowtie S| \,/\, R = R_0 \text{ et } |S| = s] = \frac{1}{\binom{d_A\, d_C}{s}} \times [y^s]\, \frac{\partial f}{\partial x}(1,y)$$

$$. \frac{\partial f}{\partial x}(x,y) = f(x,y) \times d_C . \sum_{i=1}^{d_A} \frac{j_i . y . x^{j_i - 1}}{1 + x^{j_i} y} .$$

$$\text{Donc } \frac{\partial f}{\partial x}(1,y) = d_C\, y . \left(\sum_{i=1}^{d_A} j_i \right) . (1+y)^{d_A\, d_C - 1} .$$

$$\text{Or : } \sum_{i=1}^{d_A} j_i = |R_0|$$

D'où : $\frac{\partial f}{\partial x}(1,y) = |R_0|\, d_C\, y\, (1+y)^{d_A d_C - 1}$

et $[y^s]\frac{\partial f}{\partial x}(1,y) = |R_0|\, d_C \cdot \binom{d_A\, d_C - 1}{s-1}$

Donc : $E[|R \bowtie S| / R = R_0 \text{ et } |S| = s] = \frac{|R_0| . s}{d_A}$

c) Variance

$$\sigma^2 = \sum_{t \geqslant 0} t(t-1)P[|R \bowtie S| = t / R = R_0 \text{ et } |S| = s] + E[|R \bowtie S|] - E[|R \bowtie S|]^2$$

$$\cdot \sum_{t \geqslant 0} t(t-1)P[|R \bowtie S| = t / R = R_0 \text{ et } |S| = s]$$

$$= \frac{1}{\binom{d_A\, d_C}{s}} \cdot [y^s]\frac{\partial^2 f}{\partial x^2}(1,y)$$

$$= \frac{1}{\binom{d_A\, d_C}{s}} \times \left[\binom{d_A\, d_C - 2}{s-2} d_C^2 |R_0|^2 + \left(\sum_i j_i^2\right) d_C \binom{d_A\, d_C - 2}{s-1}\right.$$

$$\left. - |R_0| d_C \binom{d_A\, d_C - 1}{s-1}\right]$$

$$= \frac{|R_0|^2 s(s-1) d_C}{d_A(d_A d_C - 1)} + \left(\sum_i j_i^2\right) \frac{s(d_A d_C - s)}{d_A(d_A d_C - 1)} - \frac{|R_0|\, s}{d_A} \cdot$$

En reportant dans l'expression de σ^2 , nous obtenons :

$$\sigma^2 = \frac{|R_0|^2 s(s-1)d_C}{d_A(d_A d_C - 1)} + \left(\sum_i j_i^2\right) \frac{s(d_A d_C - s)}{d_A(d_A d_C - 1)} - \frac{|R_0|^2 s^2}{d_A^2}$$

6) Loi de $|R \bowtie S| / |R| = r$ et $|S| = s$

La probabilité que $R \bowtie S$ soit de taille t , sachant que R est de taille r , et S de taille s , est :

$$P[|R \bowtie S| = t / |R| = r \text{ et } |S| = s]$$

$$= \frac{[x^t y^s] \sum_{R_0, tq\ |R_0| = r} \prod_{i=1}^{d_A} (1 + x^{j_i} y)^{d_C}}{\binom{d_A d_B}{r} \binom{d_A d_C}{s}}$$

j_i étant le nombre de fois où x_i apparaît dans R_0 .

La taille moyenne de $R \bowtie S$ est : $\frac{rs}{d_A}$.

Démonstration

a) $P[|R \bowtie S| = t / |R| = r \text{ et } |S| = s]$

$$= \sum_{R_0, |R_0| = r} P[|R \bowtie S| = t / R = R_0 \text{ et } |S| = s] \times P(R = R_0 / |R| = r)$$

$$\cdot\ P[R = R_0 \ /\ |R| = r] = \frac{1}{\binom{d_A\ d_B}{r}}$$

$$\cdot\ P[|R \bowtie S| = t \ /\ R = R_0 \ \text{ et } \ |S| = s]$$

$$= \frac{1}{\binom{d_A\ d_C}{s}} [x^t y^s] \prod_{i=1}^{d_A} (1 + x^{j_i} y)^{d_C} \quad .$$

Donc :

$$P[|R \bowtie S| = t \ /\ |R| = r \ \text{ et } \ |S| = s]$$

$$= \frac{1}{\binom{d_A\ d_B}{r}\binom{d_A\ d_C}{s}} \cdot \sum_{\substack{R_0,\ tq \\ |R_0| = r}} [x^t y^s] \prod_{i=1}^{d_A} (1 + x^{j_i} y)^{d_C}$$

$$= \frac{1}{\binom{d_A\ d_B}{r}\binom{d_A\ d_C}{s}} [x^t y^s] \sum_{\substack{j_1 \ldots j_{d_A} \geqslant 0 \\ j_1 + \ldots + j_{d_A} = r}} \prod_{i=1}^{d_A} (1 + x^{j_i} y)^{d_C} \quad .$$

b) Taille moyenne

Il est possible de calculer la taille moyenne par la formule $E = \sum_t t.\ P(|R \bowtie S| = t / |R| = r \ \text{ et } \ |S| = s)$. Nous donnerons ici une version plus probabiliste.

Notons, pour x_k dans D_A, j_k le nombre d'apparitions de x_k dans R, et i_k dans S :

$$|R \bowtie S| = \sum_{x_k \in D_A} j_k . i_k$$

donc : $E[|R \bowtie S| / |R| = r \text{ et } |S| = s]$

$$= \sum_{x_k \in D_A} E(j_k . i_k) \quad .$$

Les valeurs prises par R et S étant indépendantes, i_k et j_k sont aussi des variables aléatoires indépendantes. Donc $E(i_k . j_k) = E(i_k) \times E(j_k)$.

$$. \; E(i_k) = \sum_{p=0}^{r} p . Pr[x_k \text{ est } p \text{ fois dans } R]$$

$$. \; Pr[x_k \text{ est } p \text{ fois dans } R] = \frac{Np}{\binom{d_A \; d_B}{r}}$$

où N_p est le nombre de valeurs de R_1 contenant exactement p fois i_k :

$$N_k = \binom{d_B}{p} . \binom{d_A(d_B - 1)}{r - p}$$

Donc : $$E(i_k) = \sum_{p=0}^{r} p . \frac{\binom{d_B}{p}\binom{d_A(d_B - 1)}{r - p}}{\binom{d_A \; d_B}{r}}$$

$$E(i_k) = \frac{d_B}{\binom{d_A \; d_B}{r}} \sum_{p=0}^{r-1} \binom{d_B - 1}{p} \binom{d_A(d_B - 1) \; 1 - p}{r - 1 - p} .$$

Or : $$\sum_{p=0}^{r-1} \binom{d_B - 1}{p} \binom{d_A \; d_B - d_B - 1 - p}{r - 1 - p} = \binom{d_A \; d_B - 1}{r - 1} \quad .$$

Donc $$E(i_k) = \frac{d_B \cdot \binom{d_A\ d_B - 1}{r - 1}}{\binom{d_A\ d_B}{r}} = \frac{r\ d_B}{d_A d_B} = \frac{r}{d_A} \ .$$

De même, $E(j_k) = \frac{s}{d_A}$.

Nous en tirons :

$$E[|R \bowtie S| \ / \ |R| = r \ \text{ et } \ |S| = s] = d_A \times \frac{r}{d_A} \times \frac{s}{d_A} = \frac{rs}{d_A} \ .$$

c) Variance

R_0 et R_1 étant deux valeurs distinctes de R, $|R_0 \bowtie S|$ et $|R_1 \bowtie S|$ sont deux variables aléatoires indépendantes. Donc :

$$\sigma^2[|R \bowtie S| \ / \ |R| = r \ \text{ et } \ |S| = s]$$

$$= \frac{1}{\binom{d_A\ d_B}{r}} \cdot \sum_{R_0,\ \text{tq}\ |R_0| = r} \sigma^2[|R \bowtie S| \ / \ R = R_0 \ \text{ et } \ |S| = s]$$

$$= \frac{1}{\binom{d_A\ d_B}{r}} \sum_{R_0, |R_0| = r} \left[\frac{r^2 s(s-1)}{d_A(d_A\ d_C - 1)} - \frac{r^2 s^2}{{d_A}^2} + \left(\sum_{i=1}^{d_A} {j_i}^2 \right) \frac{s(d_A\ d_C - s)}{d_A(d_A\ d_C - 1)} \right]$$

$$= \frac{r^2 s(s-1)}{d_A(d_A\ d_C - 1)} - \frac{r^2 s^2}{d_A} + \frac{s(d_A\ d_C - s)}{\binom{d_A\ d_B}{r} d_A(d_A\ d_C - 1)} \cdot \sum_{R_0, |R_0| = r} \left(\sum_{i=1}^{d_A} {j_i}^2 \right)$$

B - LOIS GENERALES

1) Hypothèses

Nous supposons que la probabilité d'un tableau donné R est proportionnelle à la probabilité $P(\ell)$ de chacune de ses lignes :

$$* \quad P[R \;/\; |R| = r] = C(r) \prod_{\ell \in R} P(\ell)$$

$$\text{où} \quad C(r) = \left[\sum_{R \text{ tq } |R| = r} \prod_{\ell \in R} P(\ell) \right]^{-1}$$

* les valeurs possibles du tableau R sont données par le polynôme $P_R(x_1,\ldots,x_D) = \prod_{\ell=1}^{D} (1 + P_\ell \, x_p)$, les lignes pouvant être prises étant $x_1 \ldots x_D$, de probabilités respectives $P_1 \ldots P_D$.

La fonction génératrice des valeurs de R est donc :

$$\frac{P_R(x_1,\ldots,x_D)}{P_R(1,\ldots,1)} = f(x_1,\ldots,x_D) = \prod_{\ell=1}^{D} \left(\frac{1 + P_\ell \, x_p}{1 + P_\ell} \right)$$

* En considérant ici les schémas de relation R et S définis plus haut, nous avons les polynômes formels et fonctions génératrices suivants :

$$P_R(x_1,\ldots,x_{d_A},y_1,\ldots,y_{d_B}) = \prod_{t,s} (1 + p_t \, q_s \, x_t \, y_s)$$

$$f_R(x_1,\ldots,x_{d_A},y_1,\ldots,y_{d_B}) = \prod_{t,s} \left(\frac{1 + p_t \, q_s \, x_t \, y_s}{1 + p_t \, q_s} \right)$$

$$P_S(x_1,\ldots,x_{d_A},z_1,\ldots,z_{d_C}) = \prod_{t,u} (1 + p_t \, r_u \, x_t \, z_u)$$

$$f_S(x_1,\ldots,x_{d_A},z_1,\ldots,z_{d_C}) = \prod_{t,u}\left(\frac{1+p_t\ r_u\ x_t\ z_u}{1+p_t\ r_u}\right)$$

où

$$\begin{cases} p_t = P_{D_A}(x_t) \\ q_s = P_{D_B}(y_s) \\ r_u = P_{D_C}(z_u) \end{cases}$$

Remarque

Dans le cas où P_{D_A} , P_{D_B} et P_{D_C} sont des lois uniformes, on ne retrouve pas les hypothèses du cas étudié précédemment.

En effet, nous avions supposé que chaque valeur possible de R avait la même probabilité : $2^{-d_A d_B}$; en particulier cette probabilité était indépendante de la taille de la valeur prise par R .

Ici, au contraire, en supposant la valeur R_0 de R de taille r : sa probabilité est

$$P(R_0) = \left(\frac{1}{d_A\ d_B}\right)^r \times \left[\prod_{\ell=1}^{d_A d_B}\left(1+\frac{1}{d_A\ d_B}\right)\right]^{-1}$$

$$= \left(\frac{d_A\ d_B}{d_A\ d_B+1}\right)^{d_A d_B} \cdot (d_A\ d_B)^{-r} \ .$$

Elle dépend donc de la taille de la relation.

2) Loi de $|R \bowtie S|/R(A)$

. La fonction génératrice de $|R \bowtie S|$, sachant que $R(A) = (x_1^{j_1},\ldots,x_{d_A}^{j_{d_A}})$, est :

$$\varphi_{(j_1\ldots j_{d_A})}(x) = \prod_{t,u}\left(\frac{1+P_t\ r_u\ x^{j_t}}{1+p_t\ r_u}\right)$$

. La probabilité que $|R \bowtie S|$ soit de taille t est donc :

$$\sum_{\substack{i_1 \ldots i_{d_A} \geq 0 \\ i_1 j_1 + \ldots = r}} \prod_{s=1}^{d_A} \left[\sum_{\substack{u_1 \ldots u_{i_s} \\ \text{tous} \\ \text{distincts}}} \frac{r_{u_1}}{1 + p_s\, r_{u_1}} \times \ldots \times \frac{r_{u_{i_s}}}{1 + p_s\, r_{u_{i_s}}} \right] \times p_s^{\,i_s}$$

. La taille moyenne de $|R \bowtie S|$ est :

$$E(|R \bowtie S| / R(A)) = \sum_{t,u} \frac{p_t\, r_u}{1 + p_t\, r_u}\, j_t = |r| d_C - \sum_{t,u} \frac{j_t}{1 + p_t\, r_u}$$

. et la variance :

$$\sigma^2 = \sum_{t,u} \frac{p_t\, r_u}{(1 + p_t\, r_u)^2}\, j_t^{\,2}$$

Démonstration

a) Si $R(A) = (x_1^{\,j_1} \ldots x_{d_A}^{\,j_{d_A}})$ et $S(A) = (x_1^{\,i_1} \ldots x_{d_A}^{\,i_{d_A}})$, alors

$|R \bowtie S| = i_1 j_1 + \ldots + i_{d_A} j_{d_A}$.

Le coefficient de x^r dans $f_S(x^{j_1}, \ldots, x^{j_{d_A}}, 1, \ldots, 1)$ est alors la probabilité que le join de R et S sur A soit de taille r . D'où l'expression de la fonction génératrice $\varphi_{(j_1, \ldots, j_{d_A})}$.

b) Pour obtenir $P[|R \bowtie S| = t / R(A)]$, nous développons $\varphi(x)$ en polynôme de x :

$$\varphi(x) = \sum_{t \geq 0} P[|R \bowtie S| = t / R(A)].x^t$$

d'où le résultat.

c) $E(|R \bowtie S|/R(A)) = \varphi'_{(j_1 \ldots j_{d_A})}(1)$.

Or : $\varphi'_{(j_1 \ldots j_{d_A})}(x) = \varphi_{(j_1 \ldots j_{d_A})}(x) \times \sum_{t,u} \frac{p_t\, r_u\, j_t\, x^{j_t-1}}{1 + p_t\, r_u\, x^{j_t}}$.

Donc $\varphi'_{(j_1 \ldots j_{d_A})}(1) = 1 \times \sum_{t,u} \frac{p_t\, r_u\, j_t}{1 + p_t\, r_u}$

d) $\sigma^2 = \varphi''_{(j_1 \ldots j_{d_A})}(1) + \varphi'_{(j_1 \ldots j_{d_A})}(1) - \varphi'^2_{(j_1 \ldots j_{d_A})}(1)$.

$$\varphi''_{(j_1 \ldots j_{d_A})}(x) = \left[\sum_{t,u} \frac{p_t\, r_u\, j_t\, x^{j_t-1}}{1 + p_t\, r_u\, x^{j_t}}\right]^2 \times \varphi_{(j_1 \ldots j_{d_A})}(x)$$

$$+ \varphi_{(j_1 \ldots j_{d_A})}(x) \cdot \sum_{t,u} j_t \cdot p_t \cdot r_u \left[\frac{(j_t-1)\, x^{j_t-2}}{1 + p_t\, r_u\, x^{j_t}} - \frac{p_t\, r_u\, j_t\, x^{2j_t-2}}{(1 + p_t\, r_u\, x^{j_t})^2}\right]$$

$$\varphi''_{(j_1 \ldots j_{d_A})}(x) = \varphi_{(j_1 \ldots j_{d_A})}(x)\left\{\left[\sum_{t,u} \frac{p_t\, r_u\, j_t\, x^{j_t-1}}{(1 + p_t\, r_u\, x^{j_t})}\right]^2\right.$$

$$+ \sum_{t,u} p_t\, r_u\, j_t^2 \frac{x^{j_t-2}}{1 + p_t\, r_u\, x^{j_t}} - \sum_{t,u} p_t^2\, r_u^2 \frac{j_t^2\, x^{2j_t-2}}{(1 + p_t\, r_u\, x^{j_t})^2}$$

$$\left. - \sum_{t,u} p_t\, r_u\, j_t \frac{x^{j_t-2}}{1 + p_t\, r_u\, x^{j_t}}\right\} .$$

Donc :

$$\varphi''_{(j_1 \ldots j_{d_A})}(1) = \varphi'^2_{(j_1 \ldots j_{d_A})}(1) - \varphi'_{(j_1 \ldots j_{d_A})}(1)$$

$$+ \sum_{t,u} \left[\frac{p_t\, r_u\, j_t^2}{1 + p_t\, r_u} - \frac{p_t^2\, r_u^2\, j_t^2}{(1 + p_t\, r_u)^2} \right]$$

d'où : $$\sigma^2 = \sum_{t,u} \frac{p_t\, r_u\, j_t^2}{(1 + p_t\, r_u)^2} \cdot \blacksquare$$

3) Loi de $|R \bowtie S| \,/\, |R| = r$

La fonction génératrice de $|R \bowtie S|$, sachant que R est de taille r, est :

$$\phi(x) : \frac{[y^r]\ F(x,y)}{[y^r]\ g_R(y) \times g_S(1)}$$

avec :

$$g_R(y) = \prod_{t=1}^{d_A} \prod_{s=1}^{d_B} (1 + p_t q_s y)$$

$$g_S(y) = \prod_{t=1}^{d_A} \prod_{u=1}^{d_C} (1 + p_t r_u y)$$

$$F(x,y) = \prod_{t=1}^{d_A} \left\{ \sum_{k=0}^{d_B} \prod_{u=1}^{d_C} (1 + p_t r_u x^k y) \times [x^k] \prod_{v=1}^{d_B} (1 + p_t q_v x) \right\} .$$

La probabilité que $R \bowtie S$ soit de taille t, sachant que R est de taille r, est :

$$P[|R \bowtie S| = t \,/\, |R| = r] = \frac{[x^t y^r]\ F(x,y)}{g_S(1) \cdot [y^r] g_R(y)} .$$

La taille moyenne de $R \bowtie S$ est :

$$E = \frac{[y^{r-1}]\left\{g_R(y).\sum_{t=1}^{d_A}\left(\sum_{s=1}^{d_B}\frac{p_t q_s}{1+p_t q_s y}\right)\left(\sum_{k=1}^{d_C}\frac{p_t r_u}{1+p_t r_u}\right)\right\}}{[y^r]\, g_R(y)}$$

et la variance :

$$\sigma^2 = \frac{[y^r]\,\frac{\partial^2 F}{\partial x^2}(1,y) + \frac{\partial F}{\partial x}(1,y)}{[y^r]\, g_R(y)\quad g_S(1)} - E^2$$

Démonstration

1) $P[|R \bowtie S| = t \,/\, |R| = r]$

$$= \sum_{R(A)} P[|R \bowtie S| = t \,/\, R(A)].\, P[R(A) \,/\, R = r] .$$

Soit $\phi(x)$ la fonction génératrice de $P[|R \bowtie S| \,/\, |R|]$:

$$\phi(x) = \sum_{t \geqslant 0} P[|R \bowtie S| = t \,/\, |R| = r] \,.\, x^t$$

$$= \sum_{R(A)} \left[\sum_{t \geqslant 0} P[|R \bowtie S| = t \,/\, R(A)].x^t\right].P[R(A) \,/\, |R| = r]$$

$$= \sum_{\substack{j_1 \ldots j_{d_A} \geqslant 0 \\ j_1 + \ldots + j_{d_A} = r}} \varphi_{(j_1 \ldots j_{d_A})}(x).P[R(A) \,/\, |R| = r] .$$

$$\text{Or :}\quad P[R(A)/|R|=r] = \frac{\prod_{t=1}^{d_A} [x^{j_t}] \left\{ \prod_{V=1}^{d_B} 1 + p_t\, q_V\, x) \right\}}{[x^r] \prod_{t,V} (1 + p_t\, q_V\, x)}$$

$$\text{En effet :}\quad P[R(A)/|R|=r] = \frac{P[R(A)]}{P[|R|=r]}$$

$$\text{et}\quad . P[|R|=r] = [x^r]\, f_R(x,\ldots,x,1,\ldots,1)$$

$$. P[R(A)] = [x_1^{j_1} \ldots x_{d_A}^{j_{d_A}}]\, f_R(x_1,\ldots,x_{d_A},1,\ldots,1)$$

$$= \prod_{t=1}^{d_A} [x^{j_t}] \left\{ \prod_{V=1}^{d_B} (1 + p_t\, q_V\, x) \right\} .$$

En remplaçant $\varphi_{(j_1 \ldots j_{d_A})}$ et $P[R(A)/|R|=r]$ par leurs expressions explicites, nous obtenons donc :

$$\phi(x) = \sum_{\substack{j_1 \ldots j_{d_A} \geqslant 0 \\ j_1 + \ldots + j_{d_A} = r}} \frac{\prod_{t=1}^{d_A} \left\{ \prod_u \left(\frac{1 + p_t\, r_u\, x^{j_t}}{1 + p_t\, r_u} \right) \times [x^{j_t}] \prod_V (1 + p_t\, q_V\, x) \right\}}{[x^r] \prod_{t,V} (1 + p_t\, q_V\, x)}$$

$$\phi(x) = \frac{\sum_{\substack{j_1 \ldots j_{d_A} \geqslant 0 \\ j_1 + \ldots + j_{d_A} = r}} \prod_{t=1}^{d_A} \left\{ \prod_u (1 + p_t\, r_u\, x^{j_t}) . [x^{j_t}] \prod_V (1 + p_t\, q_V\, x) \right\}}{\prod_{t,u} (1 + p_t\, r_u) \; . \; [x^r] \prod_{t,V} (1 + p_t\, q_V\, x)} .$$

L'identité $$\sum_{\substack{j_1 \ldots j_d \geqslant 0 \\ j_1 + \ldots + j_d = r}} a_{1,j_1} \times \ldots \times a_{d,j_d} = [y^r] \prod_{t=1}^{d} \left(\sum_{k \geqslant 0} a_{t,k} \, . \, y^k \right)$$

permet d'écrire le numérateur de ϕ comme : $[y^r] \prod_{t=1}^{d} f_t(x,y)$

avec $f_t(x,y) = \sum_{k \geqslant 0} \left\{ \prod_u (1 + p_t\ r_u\ x^k) . [x^k] \prod_v (1 + p_t\ q_v\ x) \right\} . y^k$.

Soient : $\alpha_{t,k} = [x^k] \prod_u (1 + p_t\ q_u\ x)$

$\beta_{t,k} = [x^k] \prod_v (1 + p_t\ r_v\ x)$

Alors, $f_t(x,y) = \sum_{k,j} \alpha_{t,k} \cdot \beta_{t,j} \cdot y^k \cdot x^{jk}$.

$= \sum_{k \geqslant 0} \left\{ \prod_u (1 + p_t\ r_u\ x^k\ y) . [x^k] \prod_v (1 + p_t\ q_v\ x) \right\}$.

D'où le résultat.

2) Taille moyenne

$$E[|R \bowtie S|/|R| = r] = \frac{[y^r] \frac{\partial F}{\partial X} (1,y)}{[y^r] f_R(y) . g_S(1)}$$

où $F(x,y) = \prod_{t=1}^{d_A} f_t(x,y)$

Donc : $\frac{\partial F}{\partial X}(x,y) = F(x,y) \times \sum_{t=1}^{d_A} \frac{\frac{\partial f_t}{\partial X}(x,y)}{f_t(x,y)}$

et $\frac{\partial F}{\partial X}(1,y) = F(1,y) \times \sum_{t=1}^{d_A} \frac{\frac{\partial f_t}{\partial X}(1,y)}{f_t(1,y)}$.

* $f_t(x,y) = \sum_{k,j} \alpha_{t,k}\ \beta_{t,j}\ y^k . x^{jk}$

$$\Rightarrow \frac{\partial f_t}{\partial X}(1,y) = \sum_{k,j} \alpha_{t,k}\, \beta_{t,j}\; j.k.\; y^k$$

$$= \left(\sum_k k.\alpha_{t,k}\, y^k\right)\left(\sum_j j.\beta_{t,j}\right)$$

$$= y.\frac{d}{dy}\left[\prod_u (1+p_t q_u y)\right].\frac{d}{dz}\left[\prod_v (1+p_t r_v z)\right] \qquad (1)$$

$$\frac{d}{dy}\left\{\prod_u (1+p_t q_u y)\right\} = \prod_u (1+p_t q_u y) \times \sum_{s=1}^{d_B} \frac{p_t\, q_s}{1+p_t q_s y}$$

$$\frac{d}{dz}\left\{\prod_v (1+p_t r_v z)\right\}(1) = \prod_v (1+p_t\, r_v) \times \sum_{v=1}^{d_C} \frac{p_t\, r_v}{1+p_t r_v}$$

Donc :

$$\frac{\partial f_t}{\partial X}(1,y) = y.\prod_u (1+p_t q_u y).\prod_v (1+p_t r_v) \times \sum_{s=1}^{d_B} \frac{p_t\, q_s}{1+p_t q_s y} \times \sum_{v=1}^{d_C} \frac{p_t r_v}{1+p_t r_v}$$

Par ailleurs : $f_t(1,y) = \sum_{k,j} \alpha_{t,k}.\beta_{t,j}.y^k$

$$= \prod_u (1+p_t q_u y).\prod_v (1+p_t r_v)\ .$$

Donc :

$$\frac{\frac{\partial f_t}{\partial X}(1,y)}{f_t} = y.\sum_{s=1}^{d_B} \frac{p_t\, q_s}{1+p_t q_s y} \cdot \sum_{v=1}^{d_C} \frac{p_t\, r_v}{1+p_t r_v}$$

$$\cdot\ F(1,y) = \prod_t f_t(1,y)$$

$$= \prod_{t,u} (1 + p_t\ q_u\ y) \times \prod_{t,v} (1 + p_t\ r_v)$$

$$= g_R(y)\ .\ g_S(1)$$

$$\Rightarrow \quad \frac{\partial F}{\partial X}(1,y) = y.g_R(y).g_S(1).\sum_t \Big[\sum_s \frac{p_t\ q_s}{1 + p_t q_s y}\Big]\Big[\sum_v \frac{p_t\ r_v}{1 + p_t r_v}\Big] \quad .$$

Nous en tirons :

$$E[|R \bowtie S| \ /\ |R| = r]$$

$$= \frac{[y^{r-1}]\Big\{g_R(y).\displaystyle\sum_{t=1}^{d_A}\Big(\sum_{s=1}^{d_B}\frac{p_t\ q_s}{1+p_t q_s y}\Big)\Big(\sum_{v=1}^{d_C}\frac{p_t\ r_v}{1+p_t r_v}\Big)\Big\}}{[y^r]\ g_R(y)} \quad . \quad \blacksquare$$

4) Loi de $|R \bowtie S| \ /\ R = R_0$ et $|S| = s$

La probabilité que $R \bowtie S$ soit de taille t, sachant que $R = R_0$ et $|S| = s$, est : $P[|R \bowtie S| = t \ /\ R = R \text{ et } |S| = s]$

$$= \frac{[x^t\ y^s]\ f(x,y)}{[y^s]\ g_S(y)}$$

où : $f(x,y) = \displaystyle\prod_{u=1}^{d_A}\prod_{v=1}^{d_C}(1 + p_u\ r_v\ x^{j_u}\ y)$

j_u étant le nombre d'apparitions de x_u dans R_0

$$g_S(y) = \prod_{u=1}^{d_A}\prod_{v=1}^{d_C}(1 + p_u\ r_v\ y) \quad .$$

La taille moyenne de $R \bowtie S$ est :

$$E[|R \bowtie S| \ / \ R = R_0 \ \text{ et } \ |S| = s]$$

$$= |R| \ d_C - \sum_{u,v} \frac{j_u \ [y^s] \left\{ \frac{g_S(y)}{1 + p_u \ r_v \ y} \right\}}{[y^s] \ g_S(y)} \ .$$

La variance est :

$$\sigma^2 = \frac{[y^s] \ \frac{\partial^2 f}{\partial x^2} \ (1,y)}{[y^s] \ g_S(y)} + E[|R \bowtie S|] - E(|R \bowtie S|)^2$$

Démonstration

1) La fonction génératrice de S est :

$$f_S(x_1,\ldots,x_{d_A},z_1,\ldots,z_{d_C}) = \prod_{u,v} \left(\frac{1 + p_u \ r_v \ x_u \ z_v}{1 + p_u \ r_v} \right)$$

$$= \sum_S P(S).S \ .$$

Soit $f_1(x,y)$ la fonction obtenue en remplaçant, dans f_S, chaque x_u par x^{j_u}, où j_u est le nombre d'apparitions de x_u dans R_0, et chaque z_i par y. Nous obtenons :

$$f_1(x,y) = \prod_{u,v} \left(\frac{1 + p_u \ r_v \ x^{j_u} \ y}{1 + p_u \ r_v} \right)$$

$$= \sum_{S} P(S) . x^{|R_0 \times S|} . y^{|S|}$$

$$= \sum_{s,t} P[S, \text{ tel que } |S| = s \text{ et } |R_0 \bowtie S| = t]\ x^t\ y^s \quad .$$

Donc : $P[|S| = s \text{ et } |R_0 \bowtie S| = t] = [x^t\ y^s]\ f_1(x,y)$.

Par ailleurs :

$$P[|R_0 \bowtie S| = t \,/\, |S| = s] = \frac{P[|R_0 \bowtie S| = t \text{ et } |S| = s]}{P[|S| = s]}$$

et

$$P[|S| = s] = [y^s]\ f_S(y,\ldots,y,1,\ldots,1) \ ;$$

D'où : $P[|R_0 \bowtie S| = t \,/\, |S| = s]$

$$= \frac{[x^t y^s] \prod_{u,v} \left(\frac{1 + p_u\ r_v\ x^{ju}\ y}{1 + p_u\ r_v} \right)}{[y^s] \prod_{u,v} \left(\frac{1 + p_u\ r_v\ y}{1 + p_u\ r_v} \right)}$$

$$= \frac{[x^t y^s] \prod_{u,v} (1 + p_u\ r_v\ x^{ju}\ y)}{[y^s] \prod_{u,v} (1 + p_u\ r_v\ y)} \quad .$$

D'où le résultat, en notant :

$$f(x,y) = \prod_{u,v} (1 + p_u\ r_v\ x^{ju}\ y)$$

et

$$g_S(y) = f(1,y) = \prod_{u,v} (1 + p_u\ r_v\ y) \quad .$$

2) Taille moyenne

$$E[|R \bowtie S| \ / \ R = R_0 \ \text{ et } \ |S| = s]$$

$$= \frac{\sum_t t. \ [x^t y^s] \ f(x,y)}{[y^s] \ g_S(y)}$$

$$= \frac{[y^s] \ \frac{\partial f}{\partial x} (1,y)}{[y^s] \ g_S(y)}$$

$$\cdot \frac{\partial f}{\partial x} (x,y) = f(x,y) \times \sum_{u,v} \frac{j_u \cdot p_u \cdot r_v \cdot y \cdot x^{j_u - 1}}{1 + p_u \ r_v \ x^{j_u} \ y}$$

$$\Rightarrow \frac{\partial f}{\partial x} (1,y) = f(1,y) \times \sum_{u,v} \frac{j_u \cdot p_u \cdot r_v \cdot y}{1 + p_u \ r_v \ y}$$

$$= g_S(y) \sum_{u,v} j_u \left(1 - \frac{1}{1 + p_u \ r_v \ y}\right)$$

$$= |R| . d_C . g_S(y) - g_S(y) . \sum_{u,v} \frac{j_u}{1 + p_u \ r_v \ y}$$

$$= |R| . d_C . g_S(y) - \sum_{u,v} j_u . \frac{g_S(y)}{1 + p_u \ r_v \ y} \quad .$$

Donc :

$$E[|R \bowtie S| / \ R = R_0 \ \text{ et } \ |S| = s]$$

$$= |R|.d_C - \frac{\sum_{u,v} j_u.[y^s] \left\{\frac{g_S(y)}{1 + p_u\ r_v\ y}\right\}}{[y^s]\ g_S(y)} \quad .$$

3) Variance

Notons σ^2 la variance de $|R \bowtie S|$, et E sa taille moyenne, sachant que $R = R_0$ et $|S| = s$:

$$\sigma^2 = E[|R \bowtie S|^2 / R = R_0 \text{ et } |S| = s] - E^2$$

$$= \sum_t t^2\ P[|R \bowtie S| = t \ / \ R = R_0 \text{ et } |S| = s] - E^2$$

$$= \sum_t t(t-1)\ \frac{[x^t y^s]\ f(x,y)}{[y^s]\ g_S(y)} + E - E^2$$

$$= \frac{[y^s] \sum_t t(t-1)\ [x^t]\ f(x,y)}{[y^s]\ g_S(y)} + E - E^2$$

$$= \frac{[y^s]\ \frac{\partial^2 f}{\partial x^2}\ (1,y)}{[y^s]\ g_S(y)} + E - E^2 \quad . \quad \blacksquare$$

5) Loi de $|R \bowtie S| / |R|$ et $|S|$

La probabilité que $R \bowtie S$ soit de taille t, sachant que R est de taille r, et S de taille s, est :

$$\frac{[x^t y^r z^s]\ \psi(x,y,z)}{[y^r]\ g_R(y) \times [y^s]\ g_S(y)} \quad ; \text{ où}$$

$$\psi(x,y,z) = \prod_{t=1}^{d_A} \left\{ \sum_{k=0}^{d_B} \prod_{v=1}^{d_C} (1 + p_t\ r_v\ x^k\ z).[y^k] \prod_{u=1}^{d_B} (1 + p_t\ q_u\ y).y^k \right\} .$$

La taille moyenne de $R \bowtie S$ est :

$$E = \frac{[y^r z^s]\ \frac{\partial \psi}{\partial x}(1,y,z)}{[y^r]\ g_R(y).[y^s]\ g_S(y)}$$

$$= \sum_{t=1}^{d_A} \frac{[y^{r-1}]\left\{ g_R(y).\sum_{u=1}^{d_B} \frac{p_t q_u}{1+p_t q_u y} \right\}}{[y^r]\ g_R(y)} \times \frac{[z^{s-1}]\left\{ g_S(y).\sum_{v=1}^{d_C} \frac{p_t r_v}{1+p_t r_v z} \right\}}{[z^s]\ g_S(z)}$$

et la variance :

$$\sigma^2 = \frac{[y^r z^s]\ \frac{\partial^2 \psi}{\partial x^2}(1,y,z)}{[y^r]\ g_R(y).[y^s] g_S(y)} + E - E^2 .$$

Démonstration

1) Notons P(t/r et s) la probabilité

$$P(|R \bowtie S| = t \;/\; |R| = r \text{ et } |S| = s) :$$

$$P(t/r \text{ et } s) = \sum_{R(A)} P[|R \bowtie S| = t/R(A) \text{ et } |S| = s] \times P[R(A)/|R| = r] .$$

Or : . $P[|R \bowtie S| = t \;/\; R(A) \text{ et } |S| = s]$

$= P[|R \bowtie S| = t \;/\; R_0 \text{ et } |S| = s]$

pour un R_0 , tel que $R_0(A) = x_1^{j_1} \ldots x_{d_A}^{j_{d_A}}$

$$= \frac{[x^t y^s] \prod_{t,v} (1 + p_t\, r_v\, x^{j_u}\, y)}{[y^s] \prod_{t,v} (1 + p_t\, r_v\, y)}$$

$$. P[R(A)/|R| = r] = \frac{\prod_{t=1}^{d_A} [z^{j_u}] \prod_{u=1}^{d_B} (1 + p_t\, q_u\, z)}{[z^r] \prod_{t,u} (1 + p_t\, q_u\, z)} .$$

Donc :

$$P(t/r \text{ et } s) = \frac{[x^t\, y^s]\, \phi(x,y)}{[y^s] \prod_{t,v} (1 + p_t\, r_v\, y) . [z^r] \prod_{t,u} (1 + p_t\, q_u\, z)}$$

où

$$\phi(x,y) = \sum_{\substack{j_1 \ldots j_{d_A} \geqslant 0 \\ j_1 + \ldots + j_{d_A} = r}} \prod_{t=1}^{d_A} \left\{ [z^{j_t}] \prod_{u=1}^{d_B} (1 + p_t\, q_u\, z) \times \prod_{v=1}^{d_C} (1 + p_t\, r_v\, x^{j_t}\, y) \right\} .$$

Par ailleurs, la relation :

$$\sum_{\substack{j_1 \ldots j_d \geqslant 0 \\ j_1 + \ldots + j_d = r}} a_{1,j_1} \times \cdots \times a_{d,j_d} = [z^r] \prod_{u=1}^{d} \left(\sum_{k \geqslant 0} a_{u,k}\, z^k \right)$$

nous permet d'écrire $\phi(x,y)$ comme $[z^r]\psi(x,y,z)$, où :

$$\psi(x,y,z) = \prod_{u=1}^{d_A} \left(\sum_{k \geqslant 0} [z^k] \prod_{v=1}^{d_B} (1+p_u q_v z) \times \prod_{v=1}^{d_C} (1+p_u r_v x^k y) \times z^k \right)$$

$$= \prod_{u=1}^{d_A} \left(\sum_{k \geqslant 0} [t^k] \prod_{v=1}^{d_B} (1 + p_u q_v . t . z) \times \prod_{v=1}^{d_C} (1+p_u r_v x^k\, y) \right) .$$

En notant $\alpha_{t,k} = [t^k] \prod_{v=1}^{d_B} (1 + p_t\, q_v\, t)$

et $\beta_{t,k} = [t^k] \prod_{v=1}^{d_B} (1 + p_t\, r_v\, t)$,

nous obtenons :

$$\psi(x,y,z) = \prod_{t=1}^{d_A} \left[\sum_{k=0}^{d_B} \sum_{j=0}^{d_C} \alpha_{t,k}\, \beta_{t,j} \cdot x^{j_k}\, y^j\, z^k \right] .$$

2) Taille moyenne

$$E[|R \bowtie S| \ / \ r \text{ et } s] = \frac{[y^r z^s] \frac{\partial \psi}{\partial x} (1,y,z)}{[y^r] g_R(y) . [y^s] g_S(y)}$$

$$. \ \psi(x,y,z) = \prod_{t=1}^{d_A} f_t(x,y,z)$$

avec $f_t(x,y,z) = \sum_{j=0}^{d_B} \sum_{k=0}^{d_C} \alpha_{j,t}\ \beta_{k,t}\ .x^{j^k}.y^j.z^k$.

. Donc $\frac{\partial \psi}{\partial x}(x,y,z) = \psi(x,y,z) \times \sum_{t=1}^{d_A} \frac{\frac{\partial f_t}{\partial x}(x,y,z)}{f_t(x,y,z)}$

et $\frac{\partial \psi}{\partial x}(1,y,z) = \psi(1,y,z) . \sum_{t=1}^{d_A} \frac{\frac{\partial f_t}{\partial x}(1,y,z)}{f_t(1,y,z)}$

* $\psi(1,y,z) = \prod_{t=1}^{d_A} f_t(1,y,z)$

$$f_t(1,y,z) = \sum_{j,k} \alpha_{j,t}.\beta_{k,t}.y^j.z^k$$

$$= \left(\sum_j \alpha_{j,t}.y^j\right).\left(\sum_k \beta_{k,t}.z^k\right)$$

$$= \prod_{j=1}^{d_B} (1 + p_t\ q_j\ y) . \prod_{k=1}^{d_C} (1 + p_t\ r_k\ z)$$

Donc $\psi(1,y,z) = \prod_{t,j} (1 + p_t\ q_j\ y) . \prod_{t,k} (1 + p_t\ r_k\ z)$

$$= g_R(y)\quad g_S(z) .$$

* $\frac{\partial f_t}{\partial x}(x,y,z) = \sum_{j,k} (j\alpha_{t,j}\ y^j).(k\ \beta_{t,k}\ z^k).x^{j_k - 1}$

$$\Rightarrow \frac{\partial f_t}{\partial x}(1,y,z) = \left(\sum_j j\ \alpha_{t,j}\ y^j\right)\left(\sum_k k\ \beta_{t,k}\ z^k\right)$$

$$= yz \cdot \frac{d}{dy}\left[\prod_v (1+p_t q_u y)\right] \cdot \frac{d}{dz}\left[\prod_v (1+p_t r_v z)\right]$$

$$\cdot \frac{d}{dy}\left[\prod_u (1+p_t q_u y)\right] = \prod_u (1+p_t q_u y) \times \sum_u \frac{p_t q_u}{1+p_t q_u y}$$

$$\cdot \frac{d}{dz}\left[\prod_v (1+p_t r_v z)\right] = \prod_v (1+p_t r_v z) \times \sum_s \frac{p_t r_s}{1+p_t r_s y}$$

$$\text{Donc : } \frac{\partial f_t}{\partial x}(1,y,z) = f_t(1,y,z) \cdot yz \cdot \left(\sum_u \frac{p_t q_u}{1+p_t q_u y}\right)\left(\sum_v \frac{p_t r_v}{1+p_t r_v z}\right)$$

$$\text{et} \quad \frac{\partial \psi(1,y,z)}{\partial x} = \sum_{t=1}^{d_A} yz \cdot g_R(y) \cdot g_S(z) \left(\sum_u \frac{p_t q_u}{1+p_t q_u y}\right)\left(\sum_v \frac{p_t r_v}{1+p_t r_v z}\right)$$

d'où finalement :

$$E[|R \bowtie S| \,/\, r \text{ et } s]$$

$$= \frac{[y^{r-1}\ z^{s-1}] \sum_{t=1}^{d_A} g_R(y) \cdot g_S(z) \left(\sum_u \frac{p_t q_u}{1+p_t q_u y}\right)\left(\sum_v \frac{p_t r_v}{1+p_t r_v z}\right)}{[y^r]\ g_R(y) \quad [y^s]\ g_S(y)}$$

$$= \sum_{t=1}^{d_A} \frac{[y^{r-1}]\left\{g_R(y) \sum_u \frac{p_t q_u}{1+p_t q_u y}\right\}}{[y^r]\ g_R(y)} \quad \frac{[z^{s-1}]\left\{g_S(y) \sum_v \frac{p_t r_v}{1+p_t r_v z}\right\}}{[z^s]\ g_S(z)} \cdot \blacksquare$$

CHAPITRE VI

DEMIJOINTURE

A - LOIS UNIFORMES

1) Hypothèses

Nous supposons que : R et S ont pour attributs respectifs, non nécessairement simples, A et B , et A et C , où $A \cap B = \phi = A \cap C = B \cap C$.

Il n'y a pas de dépendance fonctionnelle dans R ou S .

Les valeurs prises par R et S sont indépendantes.

Les lois sur D_A , D_B et D_C sont uniformes, et les lois sur $D_A \times D_B$ et $D_A \times D_C$ sont les lois produites. (indépendance des colonnes A et B (resp. A et C)).

Nous en tirons :

Pour une taille donnée r (resp. s) , les valeurs de R (resp. S) sont équiprobables :

$$P[R = R_0 \ / \ |R_0| = r] = \frac{1}{\binom{d_A \ d_B}{r}}$$

2) Résultat

$$P[|\Pi_R(R \bowtie S)| = t \ / \ |R| = r \text{ et } |S| = s]$$

$$= \frac{1}{\binom{d_A \ d_B}{r}\binom{d_A \ d_C}{s}} \sum_{u=1}^{s} \binom{d_A \ d_B - ud_B}{r \ - \ t}\binom{ud_B}{t}\binom{d_A}{u} X_{u,s}(d_C)$$

$$\text{où} \quad X_{u,s}(d_C) = \sum_{\substack{n_1 \ldots n_u \geqslant 1 \\ n_1 + \ldots + n_u = s}} \binom{d_C}{n_1} \ldots \binom{d_C}{n_u}$$

Démonstration

1°) Supposons que S ne comporte pas d'attribut C . Alors :

$$P[|\Pi_R(R \bowtie S)| = t \;/\; |R| = r \text{ et } |S| = s] = \frac{\binom{d_A\, d_B - sd_B}{r - t}\binom{sd_B}{t}}{\binom{d_A\, d_B}{r}}$$

En effet :

La probabilité cherchée est le rapport du nombre N_C de couples (R,S) convenant au nombre N_t de couples total.

* N_t = Nombre de valeurs de R × Nombre de valeurs de S

$$= \binom{d_A\, d_B}{r}\binom{d_A}{s}$$

* Soit $T = \Pi_R[R \bowtie S]$

$T = \{(a,b) \in R \text{ , tel que } a \in S\} = R \bowtie S$

$(T \subset R)$

Le choix de (R,S), tel que $R \times S$ soit de taille t , se fait ainsi :

. Fixons les s éléments de S : $\binom{d_A}{s}$ possibilités.

. Puis, parmi les (sd_B) lignes possibles pour T, choisissons $T : \binom{sd_B}{t}$ possibilités.

. Enfin, parmi les $(d_A - s) \times d_B$ lignes de R correspondant à des éléments de A non dans S , choisissons les $(r - t)$ éléments de $R \setminus T$: $\binom{d_A\, d_B - sd_B}{r - t}$ choix.

D'où

$$N_C = \binom{d_A}{s}\binom{sd_B}{t}\binom{d_A\, d_B - sd_B}{r - t}$$

Nous en tirons :

$$P[|\Pi_R(R \bowtie S)| = t \;/\; |R| = r \text{ et } |S| = s] = \frac{\binom{sd_B}{t}\binom{d_A\, d_B - sd_B}{r - t}}{\binom{d_A\, d_B}{r}}$$

2°) Dans le cas général où la colonne C existe bien, nous remarquons que :

$$\Pi_R[R \bowtie S] = \{(a,b) \in R \text{ , tel que } a \in \Pi_A(S)\} = R \bowtie \Pi_A(S)$$

Donc, en conditionnant par la taille de $\Pi_A(S)$, nous avons :

$$P[|\Pi_R(R \bowtie S)| = t \,/\, |R| = r \text{ et } |S| = s]$$

$$= \sum_{u=1}^{s} P[|\Pi_R(R \bowtie S)| = t \,/\, |R| = r \text{ et } |\Pi_A(S)| = u] \times P[|\Pi_A(S)| = u \,/\, |S| = s]$$

$$* \; P[|\Pi_R(R \bowtie S)| = t \,/\, |R| = r \text{ et } |\Pi_A(S)| = u]$$

$$= \frac{\binom{d_A \; d_B - u d_B}{r - t} \binom{u d_B}{t}}{\binom{d_A \; d_B}{r}}$$

$$* \; P[|\Pi_A(S)| = u \,/\, |S| = s]$$

$$= \frac{\binom{d_A}{u} X_{u,s} \; (d_C)}{\binom{d_A \; d_C}{s}}$$

(cf. chapitre sur la taille des projections)

D'où le résultat. ■

Taille moyenne

La taille moyenne de $\Pi_R[R \bowtie S]$ *, sachant que* $|R| = r$ et $|S| = s$ *, est :*

$$r\left[1 - \frac{\binom{d_A\, d_C - d_C}{s}}{\binom{d_A\, d_C}{s}}\right] \simeq \frac{rs}{d_A}\left[1 - \frac{s-1}{2\, d_A\, d_C}\right]$$

Démonstration

1°) Supposons d'abord que S ait pour seule colonne A .

$$E[|\Pi_R(R \bowtie S)| \,/\, |R| = r \text{ et } |S| = s]$$

$$= \sum_{t=0}^{r} t.P[|\Pi_R(R \bowtie S)| = t \,/\, |R| = r \text{ et } |S| = s]$$

$$= \sum_{t=0}^{r} t.\frac{\binom{sd_B}{t}\binom{d_A\, d_B - sd_B}{r-t}}{\binom{d_A\, d_B}{r}}$$

$$= \frac{sd_B}{\binom{d_A\, d_B}{r}} \sum_{t=0}^{r-1} \binom{sd_B - 1}{t}\binom{d_A\, d_B - sd_B}{(r-1)-t}$$

$$= \frac{sd_B}{\binom{d_A\, d_B}{r}} \times \binom{d_A\, d_B - 1}{r-1} = \frac{rs}{d_A} \quad .$$

2°) Dans le cas général :

$$E[|\Pi_R(R \bowtie S)| \,/\, |R| = r \text{ et } |S| = s]$$

$$= \sum_{\substack{1 \leqslant u \leqslant s \\ 1 \leqslant t \leqslant r}} t.P[|\Pi_R(R \bowtie S)| = t \,/\, |R| = r \text{ et } |\Pi_A(S)| = u] \times P[|\Pi_A(S)| = u \,/\, |S| = s]$$

$$= \sum_{1 \leqslant u \leqslant s} P[|\Pi_A(S)| = u \,/\, |S| = s] \times \sum_{1 \leqslant t \leqslant r} t.P[|\Pi_R(R \bowtie S)| = t \,/\, |R| = r \text{ et } |\Pi_A(S)| = u] .$$

Le facteur de droite n'est autre que la taille moyenne du semi-join de R avec $\Pi_A(S)$, et vaut donc : $\frac{r.u}{d_A}$.

D'où :

$$E[|\Pi_R(R \bowtie S)| \,/\, |R| = r \text{ et } |S| = s]$$

$$= \frac{r}{d_A} . \sum_{1 \leqslant u \leqslant s} u.\, P[|\Pi_A(S)| = u \,/\, |S| = s]$$

$$= \frac{r}{d_A} . E[|\Pi_A(S)| \,/\, |S| = s]$$

$$= \frac{r}{d_A} \times d_A \left[1 - \frac{\binom{(d_A - 1)d_C}{s}}{\binom{d_A\, d_C}{s}} \right]$$

D'où le résultat et la formule approchée. ■

Variance

$$\sigma^2(X) = r^2[a_I^2(s) - a_I^2(s)] + \frac{r(r-1)(d_A-1)d_B}{d_A d_B - 1}\,[a_2(s) - a_1(s)]$$

$$\simeq \frac{rs}{d_A}\left(1 - \frac{3s-1}{2d_A}\right)$$

où $\quad a_1(s) = \dfrac{\binom{d_A\ d_C - d_C}{s}}{\binom{d_A\ d_C}{s}} \quad$ et $\quad a_2(s) = \dfrac{\binom{d_A\ d_C - 2d_C}{s}}{\binom{d_A\ d_C}{s}}$

Démonstration

1°) Supposons que S ait pour seul attribut A .

Notons $X = |R \bowtie S| \,/\, |R| = r$ et $|S| = s$:

$$\sigma^2(X) = E(X^2) - (E(X))^2$$

. $E(X) = \dfrac{rs}{d_A}$

. $E(X^2) = \displaystyle\sum_{t=1}^{r} t^2 P(X = t)$

$$= \sum_{t=2}^{r} t(t-1)\, P(X = t) + \sum_{t=1}^{r} tP(X = t)$$

$$= E(X) + \sum_{t=2}^{r} t(t-1)\, \frac{\binom{sd_B}{t}\binom{d_A\ d_B - sd_B}{r\ -\ t}}{\binom{d_A\ d_B}{r}}$$

$$= E(X) + \frac{sd_B(sd_B - 1)}{\binom{d_A\ d_B}{r}} \sum_{t=2}^{r} \binom{sd_B - 2}{t-2} \binom{d_A\ d_B - sd_B}{r - t}$$

$$= E(X) + \frac{sd_B(sd_B - 1)}{\binom{d_A\ d_B}{r}} \binom{d_A\ d_B - 2}{r-2}$$

$$= \frac{rs}{d_A} + \frac{r(r-1)s(sd_B - 1)}{d_A(d_A\ d_B - 1)}$$

Donc :

$$\sigma^2(X) = \frac{rs}{d_A} + \frac{r(r-1)s(sd_B - 1)}{d_A(d_A\ d_B - 1)} - \left(\frac{rs}{d_A}\right)^2$$

$$= \frac{rs}{d_A}\left[1 + \frac{(r-1)(sd_B - 1)}{d_A\ d_B - 1} - \frac{rs}{d_A}\right]$$

$$\sigma^2(X) = \frac{rs(d_A\ d_B - r)(d_A - s)}{d_A^{\,2}(d_A\ d_B - 1)} \ .$$

Dans le cas où r et s sont petits devant d_A et d_B :

$$\sigma^2(X) \simeq \frac{rs}{d_A}\left(1 - \frac{s}{d_A}\right)\left(1 - \frac{r}{d_A\ d_B}\right) \ .$$

2°) Dans le cas où S a bien deux attributs A et C :

$$\sigma^2(X) = E(X^2) - (E(X))^2$$

et $$E(X^2) = \sum_{t=1}^{r} t^2 P(X = t)$$

Or :

$$P(X = t) = \sum_{u=1}^{s} P[|\Pi_R(R \bowtie S)| = t \,/\, |R| = r \text{ et } |\Pi_A(S)| = u]$$

$$\times P[|\Pi_A(S)| = u \,/\, |S| = s] .$$

Donc :

$$E(X^2) = \sum_{t=1}^{r} \sum_{u=1}^{s} t^2 \times P[|\Pi_A(S)| = u \,/\, |S| = s]$$

$$\times P[|\Pi_R(R \bowtie S)| = t \,/\, |R| = r \text{ et } |\Pi_A(S)| = u] .$$

Notons, dans la suite du calcul, $Y = |\Pi_A(S)| \,/\, |S| = s$.

$$E(X^2) = \sum_{u=1}^{s} P(Y = u) \sum_{t=1}^{r} t^2 P[|\Pi_R(R \bowtie S)| = t \,/\, |\Pi_A(S)| = u \text{ et } |R| = r] .$$

Or : $$\sum_{t=1}^{r} t^2 P[|\Pi_R(R \bowtie S)| = t \,/\, |\Pi_A(S)| = u \text{ et } |R| = r]$$

$$= E[|\Pi_R(R \bowtie S)|^2 \,/\, |\Pi_A(S)| = u \text{ et } |R| = r]$$

$$= \frac{ru}{d_A} \left[\frac{d_A d_B - r}{d_A d_B - 1} + \frac{(r-1)d_B}{d_A d_B - 1} u \right] .$$

La dernière ligne s'obtient en appliquant la formule pour $E(X^2)$ dans le cas où S n'a que l'attribut A .

Donc :

$$E(X^2) = \frac{r(d_A\, d_B - r)}{d_A(d_A\, d_B - 1)} \sum_{u=1}^{s} u.\ P(Y = u)$$

$$+ \frac{r(r-1)d_B}{d_A(d_A\, d_B - 1)} \sum_{u=1}^{s} u^2\ P(Y = u)$$

i.e. : $$E(X^2) = \frac{r(d_A\, d_B - r)}{d_A(d_A\, d_B - 1)}\ E(Y) + \frac{r(r-1)d_B}{d_A(d_A\, d_B - 1)}\ E(Y^2)\ .$$

Par ailleurs :

$$E(X) = \frac{r}{d_A}\ E(Y)\ .$$

Donc :

$$\sigma^2(X) = \frac{r(d_A\, d_B - r)}{d_A(d_A\, d_B - 1)}\ E(Y) + \frac{r(r-1)d_B}{d_A(d_A\, d_B - 1)}\ E(Y^2) - \frac{r^2}{{d_A}^2}\ E(Y)^2\ .$$

En remplaçant $E(Y^2)$ par $\sigma^2(Y) + (E(Y))^2$, nous obtenons :

$$\sigma^2(X) = \frac{r(r-1)d_B}{d_A(d_A\, d_B - 1)}\ \sigma^2(Y) + \frac{r(d_A\, d_B - r)}{d_A(d_A\, d_B - 1)}\ E(Y)$$

$$+ \left[\frac{r(r-1)d_B}{d_A(d_A\, d_B - 1)} - \frac{r^2}{{d_A}^2}\right] (E(Y))^2$$

$$\sigma^2(X) = \frac{r(r-1)d_B}{d_A(d_A\, d_B - 1)}\ \sigma^2(Y) + \frac{r(d_A\, d_B - r)}{d_A(d_A\, d_B - 1)} \left(1 - \frac{E(Y)}{d_A}\right) .E(Y)\ .$$

Nous avons vu (ch. IV) que :

$$E(Y) = d_A \left[1 - \frac{\binom{d_A\ d_C - d_C}{s}}{\binom{d_A\ d_C}{s}}\right]$$

et

$$\sigma^2 Y = {d_A}^2 \left[\frac{\binom{d_A\ d_C - 2d_C}{s}}{\binom{d_A\ d_C}{s}} - \frac{\binom{d_A\ d_C - d_C}{s}^2}{\binom{d_A\ d_C}{s}^2}\right]$$

$$\sigma^2 Y = {d_A}^2 \left[\frac{\binom{d_A\ d_C - d_C}{s}}{\binom{d_A\ d_C}{s}} - \frac{\binom{d_A\ d_C - 2d_C}{s}}{\binom{d_A\ d_C}{s}}\right] .$$

Nous en tirons, après simplification :

$$\sigma^2(X) = r^2 \left[\frac{\binom{d_A d_C - d_C}{s}}{\binom{d_A\ d_C}{s}} - \frac{\binom{d_A d_C - d_C}{s}^2}{\binom{d_A\ d_C}{s}^2}\right]$$

$$+ \frac{r(r-1)(d_A-1)d_B}{d_A d_B - 1} \left[\frac{\binom{d_A d_C - 2d_C}{s}}{\binom{d_A\ d_C}{s}} - \frac{\binom{d_A d_C - d_C}{s}}{\binom{d_A\ d_C}{s}}\right]$$

Soit, en notant $a_1(s) = \dfrac{\binom{d_A\ d_C - d_C}{s}}{\binom{d_A\ d_C}{s}}$ et $a_2(s) = \dfrac{\binom{d_A\ d_C - 2d_C}{s}}{\binom{d_A\ d_C}{s}}$:

$$\sigma^2(X) = r^2 (a_1(s) - a_1^2(s)) + \frac{r(r-1)(d_A-1)d_B}{d_A d_B - 1} (a_2(s) - a_1(s)) . \blacksquare$$

Valeur approchée de $\sigma^2(X)$, quand r et s sont "petits"

En remplaçant $a_1(s)$ et $a_2(s)$ par leurs valeurs approchées*, nous obtenons :

$$\sigma^2(X) \simeq \frac{r^2 s}{d_A}\left(1 - \frac{3s-1}{2d_A}\right) = r(r-1)\,\frac{s}{d_A}\left(1 - \frac{3s-3}{2d_A}\right)\left(1 - \frac{1}{d_A}\right)$$

$$\simeq \frac{rs}{d_A}\left(1 - \frac{3s-1}{2d_A}\right) \quad .$$

B - LOIS GENERALES

1) Hypothèses

Ce sont les mêmes que précédemment, à la différence que les lois sur D_A , D_B et D_C ne sont plus supposées uniformes.

Nous définissons donc deux fonctions génératrices associées respectivement à R et S , f_R et f_S , par :

$$f_R = \sum_R P(R).R = \prod_{a=1}^{d_A} \prod_{b=1}^{d_B} \left(\frac{1 + p_a\, q_b\, x_a\, y_b}{1 + p_a\, q_b} \right)$$

$$f_S = \sum_S P(S).S = \prod_{a=1}^{d_A} \prod_{c=1}^{d_C} \left(\frac{1 + p_a\, r_c\, x_a\, z_c}{1 + p_a\, r_c} \right)$$

Nous étudions la loi de taille du semi-join de R par S : $\Pi_{R}^{A}(R \underset{A}{\bowtie} S) = R \underset{A}{\bowtie} \Pi_A(S)$, en cherchant sa fonction génératrice.

Nous supposerons d'abord $|R|$ et S connus, puis étendrons le résultat à $|R|$ et $|S|$ connus.

2) Loi de $|\Pi_R(R \bowtie S)|$ / $|R|$ et S

$$P[|\Pi_R(R \bowtie S)| = t \;/\; |R| = r \text{ et } S] = \frac{[x^t y^r]\, \phi_S(x,y)}{[y^r]\, \varphi(y)}$$

où : $$\varphi(y) = \prod_{a=1}^{d_A} \prod_{b=1}^{d_B} (1 + p_a\, q_b\, y)$$

$$\phi_S(x,y) = \varphi(y) \prod_{a \in \Pi_A(S)} \prod_{b=1}^{d_B} \left(\frac{1 + p_a\, q_b\, xy}{1 + p_a\, q_b\, y} \right)$$

Démonstration

Pour $a \in [1, d_A]$, notons j_a le nombre d'apparitions de x_a dans R :

$$|\Pi_R(R \bowtie S)| = |R \bowtie \Pi_A(S)| = \sum_{a=1}^{d_A} j_a . 1_{(x_a \in S)} .$$

Donc, si nous remplaçons, dans le monôme $x_{i_1} y_{i_1} \ldots x_{i_r} y_{i_r}$ correspondant à une valeur de R , chaque x_i par $x^{1_{(x_i \in S)}}$, et chaque y_i par y , nous obtenons :

$$x^{\sum_{k=1}^{r} 1_{(x_{i_k} \in S)} . j_{i_k}} . y^{|R|} = x^{|\Pi_R(R \times S)|} . y^{|R|} .$$

La même opération appliquée à f_R nous donne :

$$\sum_R P(R) . x^{|\Pi_R(R \bowtie S)|} . y^{|R|} = \prod_{a=1}^{d_A} \prod_{b=1}^{d_B} \left(\frac{1 + p_a q_b x^{1_{(x_a \in S)}} y}{1 + p_a q_b} \right) .$$

Donc : $P[|\Pi_R(R \bowtie S)| = t \text{ et } |R| = r / S]$

$$= [x^t y^r] \left\{ \prod_{a=1}^{d_A} \prod_{b=1}^{d_B} \left(\frac{1 + p_a q_b x^{1_{(x_a \in S)}} y}{1 + p_a q_b} \right) \right\} .$$

Par ailleurs : $P[|R| = r] = [y^r] \prod_{a,b} \left(\frac{1 + p_a q_b y}{1 + p_a q_b} \right)$

et $P[|\Pi_R(R \bowtie S)| = t / |R| = r \text{ et } S]$

$$= \frac{P[|\Pi_R(R \bowtie S)| = t \text{ et } |R| = r / S]}{P[|R| = r / S]}$$

Les valeurs prises par R et S étant indépendantes : $P[|R| = r / S] = P[|R| = r]$.

Donc, finalement : $P[|\Pi_R(R \bowtie S)| = t / |R| = r \text{ et } S]$

$$= \frac{[x^t y^r] \prod_{a,b} (1 + p_a q_b x^{1_{(x_a \in S)}} y}{[y^r] \prod_{a,b} (1 + p_a q_b y)} . \quad \blacksquare$$

3) Loi de $|\Pi_R(R \bowtie S)| \,/\, |R|$ et $|S|$

$$P[|\Pi_R(R \bowtie S)| = t \,/\, |R| = r \text{ et } |S| = s] = \frac{[x^t y^r z^s]\ \phi(x,y,z)}{[y^r]\varphi(y).[y^s]\ \psi(y)}$$

où $\varphi(y) = \prod_{a=1}^{d_A} \varphi_a(y)$, $\varphi_a(y) = \prod_{b=1}^{d_B} (1 + p_a\, q_b\, y)$

$$\psi(y) = \prod_{a=1}^{d_A} \psi_a(y),\ \psi_a(y) = \prod_{c=1}^{d_C} (1 + p_a\, r_c\, y)$$

$$\phi(x,y,z) = \sum_{k=1}^{s} [u^k]\left\{ \prod_{a=1}^{d_A} \left(\varphi_a(y) + u.\varphi_a(xy).(\psi_a(z) - 1)\right)\right\}$$

Démonstration

Notons $P(t/r \text{ et } s) = P[|\Pi_R(R \bowtie S)| = t \,/\, |R| = r \text{ et } |S| = s]$:

$$P[t/r \text{ et } s] = \sum_{\text{valeurs de } S} P[t/r \text{ et } S] \,.\, P[S/s]$$

$$.\ P[t/r \text{ et } S] = \frac{[x^t y^r]\ \phi_S(x,y)}{[y^r]\ \varphi(y)}$$

$$\text{donc } P[t/r \text{ et } s] = \frac{[x^t y^r]\left\{\sum_S \phi_S(x,y).P[S/s]\right\}}{[y^r]\ \varphi(y)}$$

$$\sum_S \phi_S(x,y).P[S \,/\, s]$$

$$= \sum_{\Pi_A(S)} \sum_{\substack{S,\ \text{tel que } \Pi_A(S) \\ \text{prenne une valeur} \\ \text{donnée}}} \phi_S(x,y)\ .\ P[S \,/\, s]\ .$$

Or : $\phi_S(x,y)$ ne dépend que de $\Pi_A(S)$. Donc :

$$\sum_S \phi_S(x,y)\ P[S/s] = \sum_{\Pi_A(S)} \phi_S(x,y)\Big(\sum_{\substack{S,\ \text{à}\ \Pi_A(S)\\ \text{fixé}}} P[S/s]\Big)$$

$$= \sum_{\Pi_A(S)} \phi_S(x,y).P[\Pi_A(S)/s]$$

* Calcul de $P[\Pi_A(S)/s]$:

$$f_S = \sum_S P(S).S = \prod_{a,c} \frac{1 + p_a\ r_c\ x_a\ z_c}{1 + p_a\ r_c}$$

Nous transformons, dans chaque monôme S de f_S x_a^j en $x_a^{1(x_a \in \Pi_A(s))}$, et z_c en z.

Nous obtenons :

$$\sum_S P(S).x_1^{1(x_1 \in \Pi_A(S))} \dots x_{d_A}^{1(x_{d_A} \in \Pi_A(S))}.\ z^{|S|}$$

$$= \frac{\prod_{a=1}^{d_A} (1 + x_a [\prod_{c=1}^{d_C} (1 + p_a\ r_c\ z) - 1])}{\prod_{a,c} (1 + p_a\ r_c)}$$

D'où : $P[\Pi_A(S) = \{x_{i_1} \dots x_{i_k}\}$ et $|S| = s]$

$$= [x_{i_1} \dots x_{i_k}\ z^s]\left\{ \frac{\prod_{a=1}^{d_A} (1 + x_a [\prod_{c=1}^{d_C} (1 + p_a\ r_c\ z) - 1])}{\prod_{a,c} (1 + p_a\ r_c)} \right\}$$

$$= \frac{[z^s] \left\{ \prod_{j=1}^{k} \left[\prod_{c=1}^{d_C} (1 + p_{i_j}\, r_c\, z) - 1 \right] \right\}}{\prod_{a,c} (1 + p_a\, r_c)}$$

En divisant par $P[|S| = s] = \dfrac{[z^s] \prod_{a,c} (1 + p_a\, r_c\, z)}{\prod_{a,c} (1 + p_a\, r_c)}$

nous obtenons :

$$P[\Pi_A(S) = \{x_{i_1} \ldots x_{i_k}\} \;/\; |S| = s]$$

$$= \frac{[z^s] \left\{ \prod_{j=1}^{k} \left[\prod_{c=1}^{d_C} (1 + p_{i_j}\, r_c\, z) - 1 \right] \right\}}{[z^s] \prod_{a,c} (1 + p_a\, r_c\, z)} \quad .$$

Reportons cette valeur dans $P(t/r$ et $s)$:

$$= \frac{[x^t\, y^r\, z^s] \left\{ \sum_{\Pi_A(S)} \phi_S(x,y) \prod_{x_a \in \Pi_A(S)} \left[\prod_{c=1}^{d_C} (1 + p_a\, r_c\, z) - 1 \right] \right\}}{[y^r]\, \varphi(y) \qquad [y^s] \prod_{a,c} (1 + p_a\, r_c\, y)} \quad .$$

Le polynôme au numérateur peut s'écrire :

$$\varphi(y) \times \sum_{k=1}^{s} \sum_{1 \leq i_1 < \ldots < i_k \leq d_A} \prod_{j=1}^{k} \prod_{b=1}^{d_B} \left(\frac{1 + p_{i_j}\, q_b\, xy}{1 + p_{i_j}\, q_b\, y} \right) \left[\prod_{c=1}^{d_C} (1 + p_{i_j}\, r_c\, z) - 1 \right]$$

Or : $\displaystyle\sum_{1 \leq i_1 < \ldots < i_k \leq d_A} \prod_{j=1}^{k} a_{i_k} = [t^k] \left\{ \prod_{j=1}^{d_A} (1 + a_j\, t) \right\}$.

Donc :

$$P[t/r \text{ et } s] = \frac{[x^t\, y^r\, z^s] \{\varphi(y).F(x,y,z)\}}{[y^r]\, \varphi(y).[y^s]\, \psi(y)}$$

où $\psi(y) = \prod_{a,c} (1 + p_a\, r_c\, y)$

et $$F(x,y,z) = \sum_{k=1}^{s} [t^k] \prod_{a=1}^{d_A} \left(1 + t \times \prod_{b=1}^{d_B} \left(\frac{1 + p_a\, q_b\, xy}{1 + p_a\, q_b\, y}\right) \times \left[\prod_{c=1}^{d_C} (1 + p_a\, r_c\, z) - 1\right]\right)$$

$$\varphi(y).F(x,y,z) = \sum_{k=1}^{s} [t^k]\, G(x,y,z,t)$$

où
$$\begin{cases} G(x,y,z,t) = \prod_{a=1}^{d_A} [\varphi_a(y) + t\, \varphi_a(xy)\, [\psi_a(z) - 1]] \\ \varphi_a(y) = \prod_{b} (1 + p_a\, q_b\, y) \\ \psi_a(z) = \prod_{c} (1 + p_a\, r_c\, z) \end{cases} \quad . \quad \blacksquare$$

4) Taille moyenne et variance

La taille moyenne de $\Pi_R(R \bowtie S)$, sachant que $|R| = r$ et $|S| = s$, est :

$$\frac{[y^r z^s] \left\{\frac{\partial \phi}{\partial X}\ (1,y,z)\right\}}{[y^r]\varphi(y).[y^s]\psi(y)} ,$$

la variance est :

$$\sigma^2 = \frac{[y^r z^s]\left\{\left(\frac{\partial^2 \phi}{\partial X^2} + \frac{\partial \phi}{\partial X}\right)(1,y,z)\right\}}{[y^r]\ \varphi(y).[y^s]\ \psi(y)} - \left(\frac{[y^r z^s]\ \frac{\partial \phi}{\partial X}\ (1,y,z)}{[y^r]\varphi(y).[y^s]\psi(y)}\right)^2$$

Démonstration

Il suffit de dériver, par rapport à x, la formule :

$$\frac{[y^r z^s]\ \phi(x,y,z)}{[y^r]\varphi(y)[y^s]\psi(z)} = \sum_{t \geq 0} P[|\Pi_R(R \bowtie S)| = t / |R| = r \text{ et } |S| = s).x^t \quad . \blacksquare$$

C O N C L U S I O N

RESULTATS OBTENUS

Nous avons obtenu les lois de probabilité de F(R) ou G(R,S), où R et S sont deux valeurs de relations d'un même schéma (sélection, intéraction, projection) ou de deux schémas distincts (jointures), de tailles connues, et F(respectivement G) une opération unaire (resp. binaire) de l'algèbre relationnelle. Nous avons caractérisé ces lois par des fonctions génératrices, ce qui nous permet un calcul plus aisé des moments : espérance et variance.

VALIDITE DES HYPOTHESES RETENUES

Nous avons fait des hypothèses qui nous semblaient "raisonnables" et assez simples. En pratique, on pourra par exemple ne pas connaître explicitement les domaines des attributs d'une relation, ou bien ces domaines ne seront pas finis. Il faudra donc les estimer, ce qui peut poser certains problèmes.

Par ailleurs, nous avons fait certaines hypothèses d'indépendance des valeurs de relations qui ne sont pas toujours vérifiées. En particulier, le cas où les deux fonctions considérées sont des vues d'une même relation n'est pas abordé.

DIRECTIONS FUTURES DE RECHERCHE

a)- Quelques problèmes non résolus

Nous n'avons pas résolu le problème de la taille d'une projection dans tous les cas : sans parler de relations non dans la troisième forme noemale ou dans la forme normale de Boyce-Codd, nous n'avons pas étudié le cas d'une relation avec une clé non simple, et où la projection se fait sur un sous-ensemble d'attributs comprenant des éléments de la clé.

Nous n'avons pas non plus étudié la taille d'une jointure ou demi-jointure de deux relations dont l'une au moins satisfait une dépendance fonctionnelle. Cette question pourrait sans doute être résolue assez facilement.

Un autre problème en suspens est celui d'opérations binaires dans le cas où les valeurs de relations considérées ne sont pas indépendantes, par exemple si elles proviennent de la décomposition d'une même relation.

Enfin, il pourrait être intéressant de chercher à étendre ces résultats au cas de domaines non bornés.

b)- Application pratique des résultats

Nous n'avons donné de résultats numériques que dans le cas de lois uniformes. Pour des lois générales, nous n'avons fourni qu'une expression théorique. En effet, leur calcul est beaucoup plus compliqué. Il serait intéressant de voir dans quelle mesure varient les résultats pour des lois générales, l'intérêt étant qu'ils sont plus faciles à calculer.
On peut aussi envisager un calcul pour des lois de Poisson, avec la même

question : quelles sont les lois de probabilité que nous pourrons estimer de manière approchée ?

Par ailleurs, nous avons donné des formules dans lesquelles intervenait la taille de la relation sur laquelle portait l'opération, ce qui, implicitement, la suppose connue ... Il serait intéressant de savoir combien de temps une estimation de cette taille, faite à un moment donné, reste exacte à 5 % près, par exemple. Ceci ne peut être étudié que sur des bases de données existantes. L'erreur commise en premant une taille fausse mériterait aussi une étude.

c)- Autres types de contraintes

Nous n'avons considéré comme contraintes d'intégrité que les dépendances fonctionnelles. Les dépendances multivaluées nous semblent rentrer dans le cadre plus général des dépendances "de type jointure". D'autres dépendances plus "statistiques" pourraient aussi être considérées.

d)- Fonctions de coût

L'utilisation des fonctions de coût pour comparer plusieurs écritures d'une requête peut être développée. Ce point n'a pas été abordé dans ce travail : nous nous sommes donnés ici des outils de base, qui restent à utiliser pour une optimisation concrete d'une requête.

e)- Fonctions génératrices

Il serait intéressant d'étudier la correspondance entre un type de schéma donné et la forme de la fonction génératrice associée. Nous n'avons considéré ici que deux types de schémas, et nous avons calculé leurs

fonctions génératrices. Nous pouvons chercher à définir des fonctions génératrices dans le cas de schémas plus compliqués et réciproquement voir s'il est possible de caractériser les dépendances fonctionnelles d'un schéma par la loi de probabilité des valeurs de la relation.

ANNEXE

ANNEXE

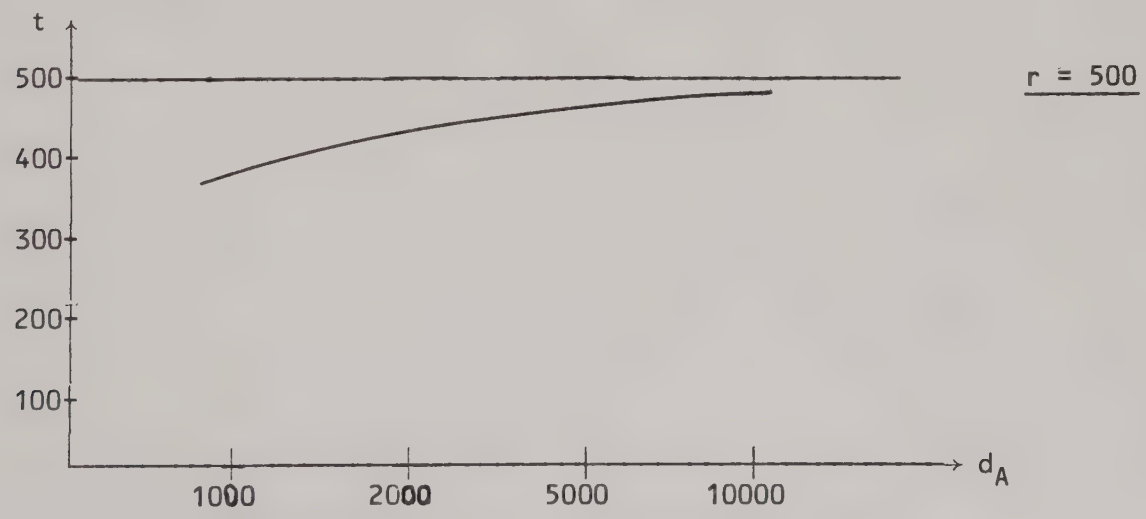
Nous donnons ici des résultats de simulations réalisées pour obtenir la taille de la projection d'une relation sans dépendance fonctionnelle, ou avec une dépendance fonctionnelle simple. Ces résultats sont en concordance avec le calcul direct de la taille moyenne.

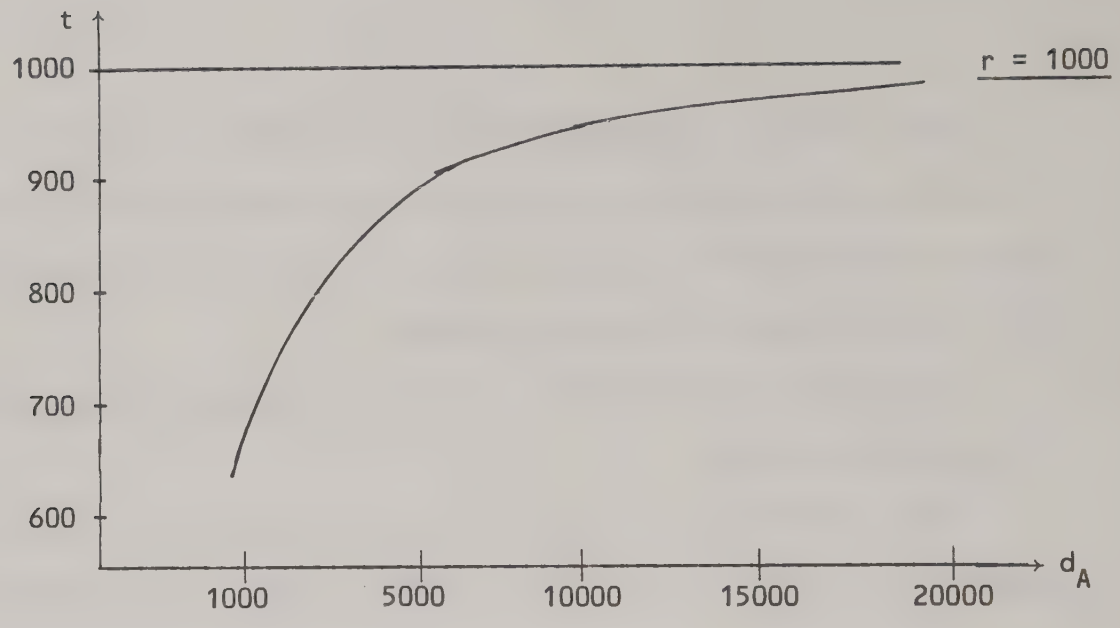
Ce travail a été réalisé sur le DEC. 20 du Centre Mondial Informatique et Ressources Humaines.

Nous y joignons un exemple d'algorithme permettant le calcul de la probabilité $P[|\Pi_A(R)| = t/|R| = r]$ dans le cas d'un schéma libre.

RELATION SANS DEPENDANCE FONCTIONNELLE

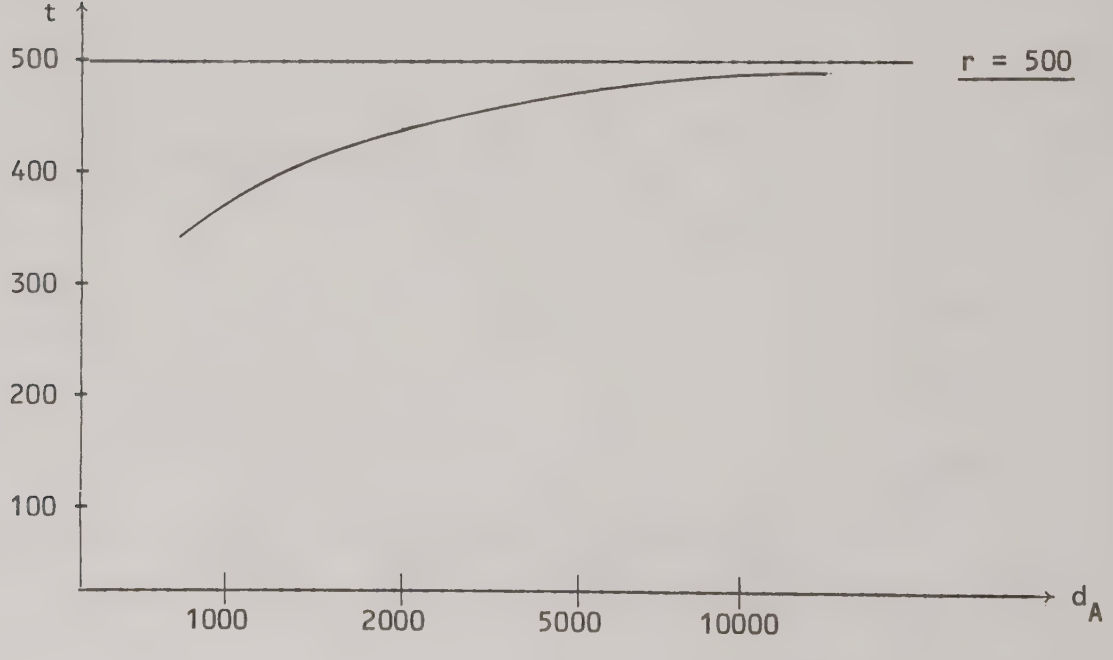
Nous étudions la taille de la projection d'une relation $R[A,B]$ sur l'attribut A. Pour une taille de relation r et une taille d_A du domaine de A fixés, la valeur de la taille moyenne subit un écart de moins de 1 % en fonction de la taille d_B du domaine de B. Nous tracerons donc les courbes $t = f(d_A)$, où t est la taille de $\Pi_A(R)$, pour $r = 500$ et $r = 1000$.

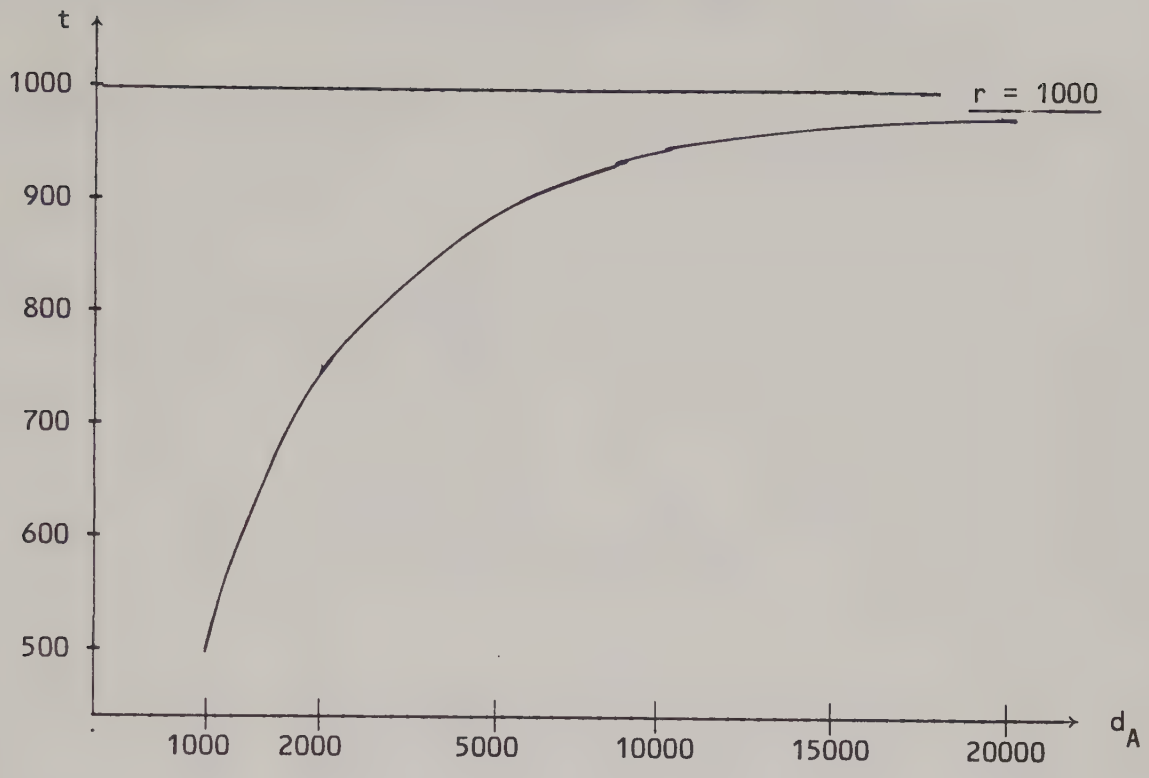




RELATION AVEC DEPENDANCE FONCTIONNELLE

Nous étudions ici la taille de la projection sur l'attribut A d'une relation R[A,B] , où B est clé. Comme ci-dessus, nous ne tenons pas compte de d_B, qui n'intervient pratiquement pas.




t
1000
900
800
700
600
500
r = 1000
1000
2000
5000
10000
15000
20000
d_A

Nous indiquons ci-dessous un algorithme permettant de calculer la probabilité :

$$P[|\pi_A(R)| = t/|R| = r] :$$

$$P[|\pi_A(R)| = t/|R| = r] = \frac{\binom{d_A}{t}}{\binom{d_A d_B}{r}} X_{r,t}(d_B)$$

où

$$X_{r,t}(d) = \sum_{\substack{n_1 \dots n_t \geq 1 \\ n_1 + \dots + n_t = r}} \binom{d}{n_1} \dots \binom{d}{n_t}$$

$$= \sum_{j=0}^{t} (-1)^{t+j} \binom{t}{j} \binom{jd}{r}$$

La première formule donnant $X_{r,t}(d)$ fait intervenir $\frac{t.r^{t-1}}{(t!)^2}$ sommes de termes $\binom{d}{n_1} \dots \binom{d}{n_t}$, et la seconde, des sommes alternées de termes pouvant être très grands. Nous donnons donc un algorithme permettant de calculer $X_{r,t}(d)$ en $t(r-t)$ étapes de façon récursive :

$$X_{r,t}(d) = \sum_{\substack{n_1 \dots n_t \geq 1 \\ n_1 + \dots + n_t = r}} \binom{d}{n_1} \dots \binom{d}{n_t}$$

$$= \sum_{n_1=1}^{r-t+1} \binom{d}{n_1} . \sum_{\substack{n_2 \dots n_t \geq 1 \\ n_2 + \dots + n_t = r - n_1}} \binom{d}{n_2} \dots \binom{d}{n_t}$$

$$= \sum_{n_1=1}^{r-t+1} \binom{d}{n_1} X_{r-n_1, t-1}(d)$$

$$= \sum_{z=t-1}^{r-1} \binom{d}{r-z} X_{z,t-1}(d) \quad .$$

Nous pouvons donc calculer récursivement $X_{r,t}(d)$, par exemple par l'algorithme suivant :

. initialiser les $X_{i,j}(d)$ à 0

. for i := 1 to (r-t) do $X_{i,1} := \binom{d}{i}$

. for j := 2 to t do

```
begin
  for i := j to (r-t+j) do
      begin
        for u := (j-1) to (i-1) do
        X[i,j] := X[i,j] + C(d, i-u) * X[u,j-1]
      end
end
```

APPENDICE

APPENDICE

I - Démonstration de : $\sum_{\substack{n_1 \ldots n_a \geq 1 \\ n_1+\ldots+n_a = b}} \binom{d}{n_1} \ldots \binom{d}{n_a} = \sum_{i=1}^{a} (-1)^{a+i} \binom{a}{i} \binom{id}{b}$

Notons $X_{a,b}(d) = \sum_{\substack{n_1 \ldots n_a \geq 1 \\ n_1+\ldots+n_a = b}} \binom{d}{n_1} \ldots \binom{d}{n_a}$

$$\sum_{\substack{n_1 \ldots n_a \geq 0 \\ n_1+\ldots+n_a = b}} \binom{d}{n_1} \ldots \binom{d}{n_a} = \binom{ad}{b} \quad (= [x^b]\{(1+x)^{ad}\})$$

L'ensemble $E = \{(n_1,\ldots,n_a)$ tels que $n_1+\ldots+n_a = b$ et $\forall i,\ n_i \geq 0\}$ peut être partitionné en $\{E_o, E_1, \ldots, E_{a-1}\}$, où chaque E_i représente l'ensemble des a-uplets dont exactement i éléments sont nuls.

Donc :

$$\binom{ad}{b} = \sum_{E} \binom{d}{n_1} \ldots \binom{d}{n_a} = \sum_{k=0}^{a-1} \sum_{E_k} \binom{d}{n_1} \ldots \binom{d}{n_a}$$

Or :

$$\sum_{E_o} \binom{d}{n_1} \ldots \binom{d}{n_a} = X_{a,b}(d)$$

et

$$\sum_{E_k} \binom{d}{n_1} \ldots \binom{d}{n_a} = \binom{a}{k} \cdot X_{a-k,b}(d) \ .\dagger$$

Nous avons donc :

$$\binom{ad}{b} = \sum_{k=0}^{a-1} \binom{a}{k} X_{a-k,b}(d)$$

d'où : $$X_{a,b}(d) = \binom{ad}{b} - \sum_{k=1}^{a-1} \binom{a}{k} X_{a-k,b}(d)$$

Nous en tirons :

$$X_{a,b}(d) = \sum_{k=1}^{a} (-1)^{a+k} \binom{a}{k} \binom{kd}{b}.$$

(la démonstration se fait sans difficulté par récurrence sur a).

Démonstration de :

$$f_{\alpha,X_1} \; 0 \; \ldots \; 0 \; f_{\alpha_1,X_n}(P) = \sum_{k=0}^{n} P_k.\alpha^k.(1-\alpha)^{n-k}$$

Notations

- $P(X_1,\ldots,X_n)$ est un polynôme à n variables
- $P_k(X)$ est le polynôme obtenu en remplaçant dans P, k des X_i par X, et les (n-k) autres par 0.
- $f_{\alpha,X_i}(P)$ est le polynôme obtenu en remplaçant chaque occurrence de X_i^n dans P par $\alpha.X^n$. Il a pour variables :

$$X_1,\ldots,X_{i-1},X,X_{i+1},\ldots,X_n.$$

Si nous abrégeons $f_{\alpha,X_i}(P)\ (X_1,\ldots)$ et $P(X_1,\ldots)$ en $f_{\alpha,X_i}(P)(X)$ et $P(X_i)$ respectivement, nous avons :

$$f_{\alpha,X_i}(P)(X) = P(0) + \alpha[P(X) - P(0)] = (1-\alpha)\ P(0) + \alpha\ P(X)) \ .$$

Démonstration

1°) Nous montrons d'abord la formule dans le cas d'une fonction de deux variables :

$$f_{\alpha,X_2}(P)\ (X_1,X) = P(X_1,0) + \alpha\,(P(X_1,X) - P(X_1,0))$$

$$= (1-\alpha)\ P(X_1,0) + \alpha\, P(X_1,X)$$

$$f_{\alpha,X_1} \circ f_{\alpha,X_2}(P)\ (X)$$

$$= (1-\alpha)\ [(1-\alpha)\ P(0,0) + \alpha P(0,X)]$$

$$+ \alpha\ [(1-\alpha)\ P(X,0) + \alpha P(X,X)]$$

$$= (1-\alpha)^2\ P(0,0) + \alpha\,(1-\alpha)\ [P(0,X) + P(X,0)] + \alpha^2 P(X,X)$$

$$= (1-\alpha)^2\ P_o(X) + \alpha\,(1-\alpha)\ P_1(X) + \alpha^2 P_2(X)$$

en posant $P_o(X) = P(0,0)$

$$P_1(X) = P(0,X) + P(X,0)$$

$$P_2(X) = P(X,X)$$

2°) Nous supposons maintenant la formule démontrée pour une fonction de (n-1) variables, et l'étendons au cas d'une fonction de n variables :

Soit $Q(X_1,\ldots,X_n) = P(X_1,\ldots,X_n)$, le polynôme à (n-1) variables obtenu en considérant X_1 comme un paramètre :

la formule de récurrence s'applique à Q ;

$$f_{\alpha,X_2} \circ \ldots\ldots \circ f_{\alpha,X_n}(Q) = \sum_{k=0}^{n-1} Q_k . \alpha^k . (1-\alpha)^{n-1-k}\ ,$$

où Q_k est défini comme plus haut.

Nous pouvons considérer Q_k comme un polynôme en X, paramétré par X_1, ou, de façon équivalente, comme un polynôme à deux variables X_1 et X, noté : $Q_k(X_1,X)$.

Les transformations f_{α,X_i}, pour $2 \leq i \leq n$, ne faisant pas intervenir X_1, nous pouvons remplacer, dans le second membre de l'équation :

$$f_{\alpha,X_1} \circ \cdots \circ f_{\alpha,X_n}(P) = f_{\alpha,X_1} \circ [f_{\alpha,X_2} \circ \cdots \circ f_{\alpha,X_n}(P)]$$

P par Q :

$$f_{\alpha,X_1} \circ \cdots \circ f_{\alpha,X_n}(P) = f_{\alpha,X_1} \circ [f_{\alpha,X_2} \circ \cdots \circ f_{\alpha_1,X_n}(Q)]$$

$$= (1-\alpha) \sum_{k=0}^{n-1} Q_k(0,X).\alpha^k.(1-\alpha)^{n-1-k}$$

$$+ \alpha. \sum_{k=0}^{n-1} Q_k(X,X).\alpha^k.(1-\alpha)^{n-1-k}$$

$$= \sum_{k=0}^{n-1} Q_k(0,X).\alpha^k.(1-\alpha)^{n-k}$$

$$+ \sum_{k=1}^{n} Q_{k-1}(X,X).\alpha^k.(1-\alpha)^{n-k}$$

$$= Q_o(0,X).(1-\alpha)^n + Q_{n-1}(X,X).\alpha^n$$

$$+ \sum_{k=1}^{n-1} [Q_k(0,X) + Q_{k-1}(X,X)]\, \alpha^k\, (1-\alpha)^{n-k}$$

- $Q_o(X_1,X) = P(X_1,0,\ldots,)$

  donc $Q_o(0,X) = P(0,0,\ldots,0) = P_o(X)$.

- $Q_{n-1}(X_1,X) = P(X_1,X,\ldots,X)$

  donc $Q_{n-1}(X,X) = P(X,-,X) = P_n(X)$

- $Q_k(0,X) + Q_{k-1}(X,X) = P_k(X)$.

En effet, $Q_k(0,X)$ donne les valeurs pour lesquelles X_1 est remplacé par 0, et k autres variables par X, et $Q_{k-1}(X,X)$ donne les valeurs pour lesquelles X_1 et (k-1) autres variables sont remplacées par X, et les (n-k) par 0.

Donc :

$$f_{\alpha,X_1} \circ \cdots \circ f_{\alpha,X_n}(P) = \sum_{k=0}^{n} P_k(X).\alpha^k \ (1-\alpha)^{n-k} \quad .$$

B I B L I O G R A P H I E

[1] A.V. AHO, Y. SAGIV, J.D. ULLMAN : "Equivalences among relational Expressions". SIAM Journal of Computing, Mai 1979.

[2] P.A. BERNSTEIN, N. GOODMAN : "Full reducers for relational queries using multiattribute semi-joins". Proceedings 1979 NBS Symposium on Computer Networks, Dec. 1979.

[3] P. BOURRET, R. DEMOLOMBE : "Choix d'une méthode de calcul de questions relationnelles base sur l'utilisation d'une fonction coût". Séminaire Bases de Données - ADI, Toulouse, Novembre 1982.

[4] P.Y. CHANG, J.M. SMITH : "Optimizing the performance of a relational algebre database interfaces". Communications of the ACM, October 1975.

[5] E.F. CODD : "A relational model of data for large shared data banks". Communications of the ACM, Juin 1970.

[6] L. COMTET : "Analyse combinatoire". Presses Universitaires de France, 1970.

[7] R. DEMOLOMBE : "Estimation of the number of t-uples satisfying a query expressed in predicate calculus language". Proceedings of VLDB, 1980.

[8] R. DEMOLOMBE : Rapport Interne CERT-ONERA, 1981.

[9] R. DEMOLOMBE : Thèse d'Etat, 1982. Université de Toulouse.

[10] W. FELLER : "An introduction to probability theory and its applications". vol. I, Wiley, 1968.

[11] P.V. HALL : "Optimization of single expressions in a relational data base system". IBM Journal of Research and Development, 1976.

[12] E. GELENBE, D. GARDY : "On the size of projections, Part. I". Information Processing Letters, vol. 14, n° 1, 1982.

[13] E. GELENBE, D. GARDY : "The size of projections of relations satisfying a functional dependancy". Proceedings of VLDB, 1982.

[14] P. RICHARD : "Evaluation of the size of a query expressed in relational algebra". Conf. ACM SIGMOD, 1981.

[15] P. RICHARD : Thèse de 3ème Cycle, 1981, Université de Paris-Sud, Orsay

[16] A.S. ROSENTHAL : "Note on the expected size of a join". ACM SIGMOD RECORD, vol. 11, n° 4, 1981.

[17] J.D. ULLMAN : "An introduction to data base theory". Computer Science Press, 1980.

[18] S. BING YAO : "Optimization of query evaluation algorithms". ACM Transactions on Data Bases, Juin 1979.

RESUME

Ce travail présente une approche probabiliste du problème de la détermination des tailles de résultats intermédiaires apparaissant lors de l'exécution d'une requête sur une base de données relationnelle. Nous supposons connues différentes lois de probabilité sur les attributs des relations initiales. Nous en déduisons dans un premier temps les lois de probabilité des valeurs possibles pour des types de relation déterminés, puis les fonctions génératrices donnant les tailles des relations obtenues par application des opérateurs de l'algèbre relationnelle : union ou intersection, sélection, projection, équijointure, auxquels nous avons ajouté la demijointure.

MOTS CLES

Modélisation
Relationnel (Modèle, Algèbre)
Taille (Estimation)
Fonction génératrice.